The Dissertation Committee for Xiaokang Shi
certifies that this is the approved version of the following dissertation:

Modeling and Optimization to Connect Layout with Silicon for Nanoscale IC

Committee:

Zhigang Pan, Supervisor

Jacob Abraham

Michael Orshansky

Lorenzo Garlappi

Ying (Frank) Liu

Modeling and Optimization to Connect Layout with Silicon for Nanoscale IC

by

Xiaokang Shi, B.S., M.S., M.S.E.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

December 2009

To my parents and friends.

Acknowledgments

It is often said that a master student's work is to give a project the best implementation, while PhD training enlightens one's insight and intelligence to explore intriguing and valuable research topics. As fruit of my PhD work, the present dissertation should not have been accomplished without many people's generous help.

First of all, I would like to thank my advisor Prof. Zhigang Pan for his supervising. His continuous supports helped me enjoy the balance of research and life. The freedom he gave me in my research allows me to investigate various topics of IC design and to accumulate experience from different approaches. I am confident that my Ph.D. experience with him would brighten my future career.

I wish to express gratitude to my Ph.D. committee members Prof. Jacob A. Abraham, Prof. Michael Orshansky, Prof. Lorenzo Garlappi and Dr. Ying (Frank) Liu, for their insightful discussions and comments despite of busy schedules.

I would like to also appreciate all help and advice from people outside UT on many issues, from technical discussions to career planning. They are Dr. Anirudh Devgan, Prof. Jiang Hu, Prof. Chris C. N. Chu, Dr. Ruiqi Tian, Prof. Jason Cong, Prof. Mark McDermott, and Dr. Rajesh Gupta.

I pride myself on my several internships at IBM T. J. Watson Research Center (Yorktown Heights, NY) and IBM Austin Research Lab. Benefited from working there I got to be exposed to cutting-edge technologies in various engineering areas (including manufacturing processing, VLSI design, CAD and areas beyond semiconductor). It is my great honor to affront research challenges in close collaboration with world-class experts and smart researchers in VLSI/CAD including Dr. David Kung, Dr. Kevin Nowka, Dr. Sani Nassif, Dr. Fook-Luen Heng, Dr. Ying (Frank) Liu, Dr. Chandramouli Visweswariah, Dr. Jin-Fuw Lee, Dr. Xiaoping Tang, Dr. Hua Xiang, Dr. Kanak Agarwal, Dr. Damir Jamsek, Dr. Haifeng Qian, Dr. Vladimir Zolotov, Dr. Zhuo Li, Dr. Cliff Sze, Dr. James Ma, Dr. Jinjun Xiong and many more. I also want to thank Blaze-DFM to give me the chance to have internship experience at the startup in Silicon Valley. My experience over there showed me the gap (or even abyss) between research in lab and practice in industry, and taught me that the devil is in the details.

In UT-Austin, I have met a lot of wonderful friends from ECE and other departments. To name a few, Tao Luo, Peng Yu, Gang Xu, Anand Ramalingam, Kun Yuan, Duo Ding, Katrina Lu, Haoxing Ren, Ashutosh Chakraborty, Suhail Ahmed, Anand Rajaram, Joydeep Mitra, James Ban, Anurag Kumar, Yilin Zhang, Wooyoung Jang, Jae-seok Yang, Kiwoon Kim, Donnie Chen, Shanhu Shen, Wei-Shen Wang, and Bin Zhang for their help and friendship. I also would like to thank Qing Ai for proofreading this dissertation. And what shall I more say? For the time would fail me to tell of Lan

Zhang, of Ning Tan, of Daifeng Wang, of Danhua Shao, of Youzhong Guo, of Jery Xuan and of others.

It is my honor to dedicate this dissertation to my family and friends back in China, especially to my parents. All the work here would not have been possible as without their love, sacrifice and support. Their companionship in both spirit and daily life has been the constant source of inspiration and impetus for me to adventure during the financial tsunami since 2008 as well as my job hunting simultaneously.

Modeling and Optimization to Connect Layout with Silicon for Nanoscale IC

Publication No. _____

Xiaokang Shi, Ph.D.
The University of Texas at Austin, 2009

Supervisor: Zhigang Pan

With continuous and aggressive scaling in semiconductor technology, there is an increasing gap between design expectation and manufactured silicon data. Research on DFM (Design for manufacturability), MFD (Manufacturing for Design) and statistical analysis have been investigated in recent years to bridge design and manufacturing. Fundamentally, layout is the final output from the design side and the input to the manufacturing side. It is also the last chance to dramatically modify the design efficiently and economically.

In this dissertation, I present the modeling and optimization work on bridging the gap between design expectation and reality, improving performance and enhancing manufacturing yield. I investigate several stages of semiconductor design development including manufacturing process, device, interconnect, and circuit level.

In the manufacturing process stage, a novel inverse lithography technology (ILT) is proposed for sub-wavelength lithography resolution enhancement. New intuitive transformations enable the method to gradually converge to the optimal solution. A highly efficient method for gradient calculation is derived based on partially coherent optical models. Dose variation is considered within the ILO process with the min-max optimization method and the computation overhead on dose process variation could be omitted. The methods are implemented in state-of-the-art industrial $32nm$ lithography environment.

After the work in the lithography process stage provides both mask optimization and post-layout silicon image simulation, my work on the first non-rectangular device modeling card extends the post-layout lithography to post-litho electrical calibration. Based on the lithography simulation results, the non-rectangular gate shapes are extracted and their effect is investigated by the proposed non-rectangular device modeling card and post-litho circuit simulation flow. This work is not only the first non-rectangular device modeling card but also compatible with industry standard device models and the parameter extraction flow.

Interconnect plays a more critical role in the nanometer scale IC design especially because of its impact on delay. The scattering effect that occurs in nanoscale wires is modeled and different methods of wire sizing/shaping are discussed. Based on closed-form resistivity model for nanometer scale Cu interconnect, new interconnect delay model and wire sizing/shaping strategies are developed.

Based on the advanced modeling of process, device and interconnect, circuit level investigation is focused on statistical timing analysis with a new latch delay model. For the first time, both combinational logic and clock distribution circuits are integrated together through statistical timing of latch outputs.

This dissertation studies the new phenomena of nanometer scale IC design and manufacture. Starting from the designed layout, through modeling, optimization and simulation, the work moves ahead to the mask pattern and silicon image, calibrates electrical properties of devices as well as circuits. Through above process, we can better connect layout with silicon data to reach design and manufacturing closure.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 The Gap between Layout and Silicon

Semiconductor technology has driven the innovation of our society for 50 years and scaling down following Moore's Law [7] has shown significant advantage on performance improvement and cost reduction. As semiconductor and IC industry is getting more and more mature, the division of labor is becoming increasing fine. For the process and manufacturing aspect, the IC development is clearly classified to different levels (as shown in Fig. 1.1). Most of the engineers only need specialization in the skills and knowledge of their own levels with a handful of components, which greatly speed up the development of the industry [1].

Furthermore, such division and specialization are not only acting onto the knowledge structures of workers in this industry, but also leading the reorganization of the commercial structure as semiconductor industry is becoming capital-intensive, more and more traditional design and manufacturing companies such as AMD, TI and so on are becoming more like fabless. The scaling of IC feature size driven by Moore's law is less and less financially affordable for semiconductor manufacturers. Moore is more (in term of investment); Moore

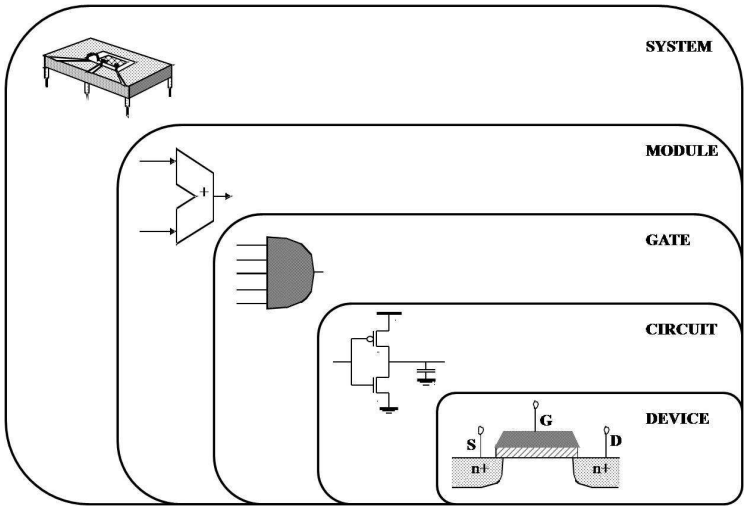


Figure 1.1: Design abstraction levels in digital circuits [1].

is less [8] (investors). During this process, there is obviously a larger and larger gap between design side and IC manufacturing side.

The final output of design side is mainly IC layout in electronic files, while IC manufacturers convert the layout to realistic silicon products. Before taking the layout out of design side, it is always fast and economic to tune the design as very this is virtual and stored in computers. However, once IC layout is taken to manufacturers, mask would be produced for the specific layout and IC products would be manufactured. It becomes much more expensive to tune the original layout design to fix the errors. However, with the feature size's scaling down, the manufactured products are more difficult to match their original design target. Fig. 1.2 shows the longer silicon validation during scaling as the gap between design target and silicon data is getting bigger.

At the same time, when the semiconductor technology moves to nanome-

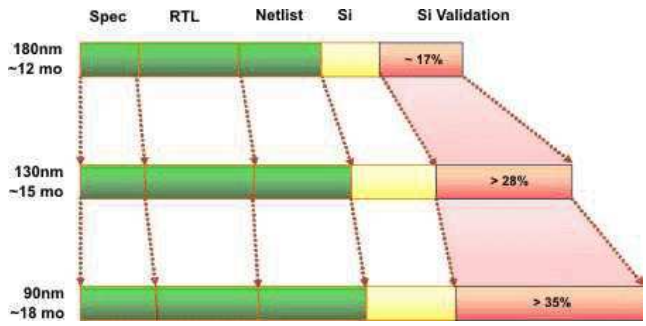


Figure 1.2: Silicon validation takes longer time during scaling [2].
If the green part can be regarded as design side and yellow is the manufacturing side, the increasing pink bars show the larger and larger gap between layout design and silicon data.

ter scale with continuous aggressive scaling, lots of new physical effects appear and challenge the traditional abstraction and approximation. For example, when the interconnect wires are scaled to nanometers or tens of nanometers, the Mesoscopic scale effect will cause scattering effect and traditional constant metal resistivity discovered by Georg Simon Ohm during 1825-30 is not valid any more. New models of interconnect delay and wire optimization methods are necessary to be developed to close the gap between estimation at layout level and measurement of silicon data.

Besides the new physical effects at nanometer scale, process variations and new manufacturing process are needed to be modeled and optimized. Since $0.25\mu m$, the patterns on wafers are manufactured at a sub-wavelength scale. Especially, after moving to $130nm$ in 2000, the lithographic wavelength has been fixed at $193nm$ while the critical dimension of silicon patterns has been scaled to $45nm$ after four generations. And it is quite possible that for the next

two coming technology nodes of $32nm$ and $22nm$, the lithographic wavelength will be still $193nm$. Traditional resolution enhancement technologies (RET) such as OPC (optical proximity correction) and phase shift mask (PSM) are enough, more advanced resolution enhancement technologies including double pattern, inverse lithography technology (ILT) and source mask optimization (SMO) would be critical to push the feature size to smaller scale. During this processing technology improvement, process variations become more and more significant.

As the patterns on layout are no longer the silicon image on wafer, traditional estimation methods based on layout will lead to larger and larger errors from the post-silicon results. Perfect rectangular poly gates on layout could be non-rectangular gates on wafers and traditional compact device models developed for rectangular gates are not valid any more. New techniques such as stressing strain also induce that the device performance depends on the layout much more strongly than traditional ways. New models to analyze this stressing effect and new optimization approach to take full advantages of stressing effect are important to link the IC layout with silicon data more accurately and efficiently [9].

Besides the devices and interconnect level challenge from the bigger gap between layout and silicon, circuit levels are also greatly impacted. For example, statistical static timing analysis and optimization are investigated a lot in recent years. Most of these works are focused on standard cells and combinational logic network. There is also some research on timing analysis

of clock distribution network with the consideration of process variations of nanometer scale IC. However, there are very few works to integrate the statistical timing analysis of both combinational logic network and clock network together. One of the key reasons is because traditional sequential cells were not well modeled for statistical timing analysis.

In this dissertation, I present new modeling and optimization techniques on bridging the gap between design expectation and reality for performance and manufacturability improvement at different stages of semiconductor technologies including manufacturing process, device, interconnect,h and circuit level.

1.2 Overview and Contributions of This Dissertation

In the manufacturing process stage, a novel inverse lithography technology (ILT) is proposed for sub-wavelength lithography resolution enhancement. It is computationally efficient and practically compatible to industry standard as well. New intuitive transformations enable the method to gradually converge to the optimal solution through a highly efficient gradient calculation derived based on partially coherent optical models. In addition to developing ILT for nominal process conditions, dose process variation aware ILO is also developed with min-max optimization and limited computation overhead. The work is verified by the newest industry $32nm$ technology. It also shows great potential to push the most front edge technology by one generation and use $32nm$ lithography technology to manufacture $22nm$ IC design.

The non-rectangular device modeling card extends the post-layout silicon image to post-litho electrical calibration. Based on the lithography simulation results, the non-rectangular gate shapes are extracted and their effect is investigated by the proposed non-rectangular device modeling card and post-litho circuit simulation flow. My work is not only the first non-rectangular device modeling card but also has the advantage of compatibility with industry standard device models and the parameter extraction flow.

Interconnect is playing a more critical role in the nanometer scale IC design especially on delay. The scattering effect of nanoscale wires is modeled for the interconnecting plan and different methods of wire sizing/shaping are discussed.

Based on the advanced modeling of process, device and interconnect, circuit level investigation is focused on statistical timing analysis with a new latch delay model. For the first time, process variations effects on timing of both combinational logic and clock distribution circuits can be integrated together through the statistical timing of latch outputs.

This dissertation discusses the emerging technologies and new phenomena of nanometer scale IC. Starting from the designed layout, by modeling, optimization and simulation, the work moves ahead to mask and silicon image, reevaluates devices as well as circuits and connect layout better with silicon data to reach design and manufacturing closure.

Chapter 2

Inverse Lithography Optimization for L2S

Optical lithography as one of the most critical process steps from designed layout to manufactured silicon, remains its wavelength at $193nm$ while the half pitch in semiconductor process technology is being pushed into tens of nanometers. Although they have been successful in previous generations of technology, traditional resolution enhancement techniques (such as OPC) cannot guarantee the optimality of the mask under increasing manufacturing variations.

In this chapter, novel inverse lithography optimization (ILO) methods will be proposed to better connect layout to silicon (L2S). Two transformations are used to convert the original integer nonlinear programming problem into an unconstrained continuous optimization problem. The min-max optimization technique is applied to ILO with dose variations. A series of highly efficient methods to calculate the necessary gradients at both the normal conditions and variable dosage conditions are also derived. The algorithms have been implemented in a state-of-the-art industrial $32nm$ lithography environment. Experimental results show that the proposed method is not only very effective for the current technology node, but also capable of achieving good resolution

for the next technology node without modifying the lithography environment. Based on the proposed inverse lithography technologies (ILT), the gap between design layout and manufactured silicon image would be reduced with not only more computation efficiency but also better tolerance to process variations.

2.1 Introduction of Lithography and Resolution Enhancement

2.1.1 Scaling and Resolution Enhancement

Optical lithography is a crucial patterning step in semiconductor manufacturing process. The images of the patterned photo-mask are projected through the high-precision optical system onto the wafer surface, which is coated with light-sensitive chemical compound, such as photo-resist. The patterns are then formed on the wafer surface after complex chemical reactions and follow-on manufacturing steps, such as development, etching and deposition [10, 11].

The optical lithography system's resolution (R) can be described by the Rayleigh Criterion:

$$R = k_1 \frac{\lambda}{NA} \tag{2.1}$$

in which λ is the wavelength of the light source, NA is the numerical aperture and k_1 is an illumination and mask pattern dependent factor which could be manipulated by the resolution enhancement techniques. Table 2.2 shows the trend of key lithography parameters in the Rayleigh Criterion in past 30 years from 1980 to 2008. If the parameters are normalized based on the data in 1997

Table 2.1: Trend of key lithography parameters.

Year	$\lambda\,(nm)$	NA	fluid n	k_1	$W_{min}\,(nm)$	DoF (nm)
1980	436	0.166	1	0.8	2101	15713
1984	436	0.28	1	0.75	1168	5450
1988	254	0.166	1	0.6	918	9754
1989	365	0.45	1	0.7	568	1706
1992	365	0.57	1	0.65	416	1023
1993	254	0.5	1	0.6	305	948
1996	365	0.63	1	0.65	377	817
1997	248	0.6	1	0.55	227	620
1999	193	0.5	1	0.6	232	720
2001	193	0.75	1	0.4	103	285
2001	365	0.65	1	0.6	337	760
2002	248	0.8	1	0.45	140	310
2004	193	0.85	1	0.35	79	204
2005	193	0.93	1.44	0.3	62	283
2005	193	0.93	1	0.3	62	153
2006	193	1.15	1.44	0.3	50	168
2008	193	1.3	1.44	0.3	45	118

(about ten years ago), the trend is shown in Fig. 2.1.

Traditionally as resolution R is linear to λ, the feature sizes were scaling down with smaller wavelengths. However when the VLSI technology pushes further into nanometer scale, the feasible wavelength of the photo-lithographic system remains unchanged at $193nm$ due to technical challenge and cost limitation. Although there have been repeatedly anticipated that extreme ultraviolet lithography (EUVL) with the wavelength of $13nm$ will replace traditional optical lithography, the availability of EUVL remains uncertain due to technical challenges and cost issues [12]. On the other hand, the physical limit of dry

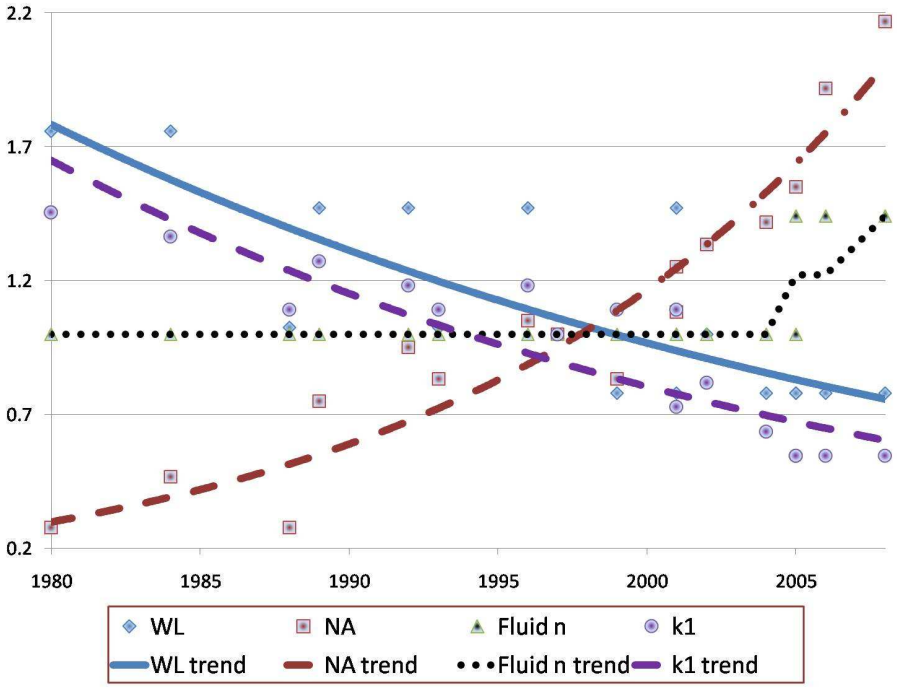


Figure 2.1: Trend of key lithography parameters for scaling down from 1980-2008 normalized with data in 1997.

lithography of NA is 1.0. The recently introduced immersion lithography has bigger NA value (1.3) by adopting fluid with high refraction index (n=1.44), but it is much harder to achieve higher NA values. Thus it is commonly recognized that k_1 remains a cost effective knob to achieve finer resolution.

2.1.2 RET, ILT and SMO

As the unavoidable diffraction become more serious when the gap between the required sub-wavelength feature size and lithography wavelength gets bigger, the final wafer images are greatly different from the patterns on

the mask. In the past few years, resolution enhancement techniques (RETs) [13] have become necessary in order to achieve the required pattern density with higher resolution and smaller k_1 in Eq. 2.1. One well-known RET is the optical proximity correction (OPC) [14–16], in which the mask pattern edges are intentionally shifted and distorted so that the desired image can be formed on the wafer. Other commonly used RETs are sub-wavelength resolution assist features (SRAF) [17, 18] and phase-shift masks (PSM) [19].

On the other hand, this particular problem can be considered from another perspective. Rather than locally perturbing the mask pattern based on layout pattern to offset the loss, the mask pattern can be treated as the input to the optical system, and the silicon image as the output. Our task then becomes how to "design" a mask to form a desired wafer image which is close enough to the favorite layout. This concept has been proposed for over 20 years [20, 21]. Whereas, the problem itself is often ill-posed in the sense that more than one input can generate the same output. For modern lithography system, it can be shown that the size of the search space is well over $2^{1,000,000}$, which is even larger than the number of atoms in the observable universe. There were some early attempts to find a feasible solution to this problem by using methods such as simulated annealing [22], genetic algorithms [23] and random pixel flipping [24]. In recent years, the growing challenges facing sub-wavelength lithography and the ever increasing complexity of traditional RETs have made this idea more attractive, which is often referred as "inverse lithography" or "computational lithography". For example, the feasibility of

inverse lithography was demonstrated in [25]. A gradient search based method was proposed in [26] to overcome the excessive computational cost. However, the cyclic transformation and non-coherence assumption in [26] are hard to manipulate and non-realistic as it is well-recognized that partially coherent models are the only acceptable model for the practical optical lithography system. Furthermore, as the technology is further pushed, manufacturing variations (e.g., dose and focus variations during the lithograph steps) have to be taken into consideration. It is challenging to systematically incorporate the process variations into the traditional RETs.

More advanced lithography optimization such as source and mask co-optimization (SMO) is presented in [27], which benefits from a new freedom of light source pattern to be optimized simultaneously during adjusting mask pattern for high quality silicon image even under process variations. Nevertheless, there is only one light source in the lithography system but millions of mask pattern windows to be optimized. For the runtime limitation, the simultaneous optimization would only be reasonable to be applied to the most critical mask pattern window and light source pattern. For example, SRAM occupied a large portion of the chip area, SMO should be applied to decide the best source pattern and SRAM pattern on mask. However, mask optimization is only possible for the rest of millions of pattern windows as the light source is greatly fixed. SMO might greatly benefit from new mask optimization method with higher computation efficiency.

All in all, more advanced resolution enhancement technologies always

require larger computation effort. The considerable amount of computing power has to be dedicated to the lithography computation including simulation, optimization and verification. For example, Mentor Graphics has used multi-processing and multi-threading on workstation clusters [28]. Specific hardware-accelerated platform for lithography computation has been used by Brion and ASML [29].

2.1.3 Background of Lithography Computation

A lithography system for semiconductor manufacturing is very complicated, and it is typically made up of four parts: illumination, mask, exposure and wafer [30]. The typically forward lithograph simulation flow is mainly made up by two steps: optical model to calculated the light intensity projected onto wafer and photo-resist etching model to convert intensity into silicon images on the wafer.

2.1.3.1 Optical model

For the sub-wavelength lithography system is very complicated as the light through the different spots on the mask would interfere with each other. Hopkins Equation [31, 32] are generally used for describing the light on top of the photo resist plane.

$$I(\mathbf{k}) = \iint_{-\infty}^{+\infty} T(\mathbf{k}+\mathbf{k}', \mathbf{k}')\, M(\mathbf{k}+\mathbf{k}')\, M^*(\mathbf{k}')\, d^2\mathbf{k}' \tag{2.2}$$

where, $I(\mathbf{k})$ is the intensity in the frequency domain. $M(\mathbf{k})$ is the mask transmission function $M(\mathbf{r})$ in the frequency domain, when $\mathbf{k} = (\mathbf{k_x}, \mathbf{k_y})$ denotes a point in the frequency domain and $\mathbf{r} = (\mathbf{x}, \mathbf{y})$ denotes a point in the spatial domain. The superscript * denotes the complex conjugation operation. $T(\mathbf{k}', \mathbf{k}'')$ is the transmission cross coefficient (TCC), given by

$$T(\mathbf{k}', \mathbf{k}'') = \iint_{-\infty}^{+\infty} J(\mathbf{k}) K(\mathbf{k} + \mathbf{k}') K^*(\mathbf{k} + \mathbf{k}'') d^2\mathbf{k} \tag{2.3}$$

where $J(\mathbf{k})$ is the illumination function and $K(\mathbf{k})$ is the projection system transfer function. TCC is a four dimensional Hermitian and positive-definite matrix after discretization. Even though there are analytical forms for above functions, the practical computation of TCC is much more complicated as other effects such as focus blurring. In the practical approach, TCC just numerically models the light intensity transmission with complicated physical phenomenon as well as effect of parameter extraction. For a given set of lithography process parameters, TCC will be calculated once and reused repeatly to calculate light intensity (such as Eq. 2.2).

Singular value decomposition (SVD) is used to truncate the size of TCC and simplify the computation of intensity [33].

$$\begin{aligned} T(\mathbf{k}', \mathbf{k}'') &= \sum_{k=1}^{(N_T \times N_T)} \sigma_k H_k(\mathbf{k}') H_k{}^*(\mathbf{k}'') \\ &\approx \sum_{k=1}^{n_K} \sigma_k H_k(\mathbf{k}') H_k{}^*(\mathbf{k}'') \end{aligned} \tag{2.4}$$

where N_T is the grid number on one dimension of TCC matrix. σ_k are the eigenvalues and eigenvectors $H_k(\mathbf{k}')$ are kernels. n_K is the number of key

components (kernels) to be stored after SVD.

After converting back to spatial domain, the intensity is:

$$\begin{aligned} I(\mathbf{r}) &= \sum_{k=1}^{(N_T \times N_T)} \sigma_k |H_k(\mathbf{r}) \otimes F(\mathbf{r})|^2 \\ &\approx \sum_{k=1}^{n_K} \sigma_k |H_k(\mathbf{r}) \otimes F(\mathbf{r})|^2 \end{aligned} \tag{2.5}$$

Through discretization, for pixel j on the wafer window, its intensity I_j is:

$$I_j = \sum_{k=1}^{n_K} \sigma_k \left| \sum_{i=1}^{MN} h_{ijk} m_i \right|^2 \tag{2.6}$$

where $h_{*,*,k}$ is the kernel k. As TCC is a four dimensional matrix after discretization, each kernel is a two dimensional matrix. And the eigenvalue σ_k indicates the weight of the corresponding kernel k.

After partially coherent system (Eq. 2.5) is presented, optical models of complete coherence (Eq. 2.7) and the non-coherence system (Eq. 2.8) are given as below. They can be regarded as certain degrees of simplification and approximation of partial coherence. The completely coherent optical system can be regarded a simple partially coherent system with only one kernel. Then the non-coherence (or incoherent) optical system assumes that the intensity of interfering lights is just the linear superposition of all intensity, which omits the phase properties of the light wave. However the wave-particle duality of light indicates the phase position is very critical for light to be wave. From the ILO aspect, optimization based on incoherent or completely coherent assumption always induces a better optimal result than partially coherent system with

many kernels does. However, after applying the more accurate partial coherence to the so-called optimal solution drawn from non-coherence or completely coherent simplification, the more realistic result will be very different from its original value and also far away from the optimal solution.

$$I(\mathbf{r}) = |Q(\mathbf{r}) \otimes F(\mathbf{r})|^2 \tag{2.7}$$

$$I(\mathbf{r}) = |Q(\mathbf{r})|^2 \otimes |F(\mathbf{r})|^2 \tag{2.8}$$

2.1.3.2 Photo Resist Model

Analytical models (such as variable threshold resist (VTR) [34] and constant threshold resist (CTR) [35]) are more widely used in practical lithography computation, though highly sophisticated photo-resist models are recommended to quantitatively describe the chemical reactions on the wafer surface [11]. Generally, computation effort on photo resist modeling can be omitted in the term of intensity calculation with the partially coherent optical model, and CTR as shown in Eq. 2.9 has acceptable accuracy for most cases [36].

$$z = \begin{cases} 1, & \text{if } I \geq t_r \\ 0, & \text{if } I < t_r \end{cases} \tag{2.9}$$

2.1.3.3 Dose and Defocus Variation

Dose and defocus variations are two key variations during lithography manufacturing process. Dose variation describes the variation from light source dosage and can modeled by modifying the light intensity on the wafer or the threshold of photo resist. Defocus variation for depth of focus (DOF) shifting

is due to wafer non-flatness, auto-focus errors, lens aberrations, focus drift, stage errors, laser wavelength drift, reticle non-flatness and etc.

Even though dose and defocus separately only have two corners for minimum and maximum values, the worst cases of pattern edges at different locations on the silicon images would depend on the different combinations of defocus and dose variation value. Thus, different from the popular corner based timing analysis with 4 combination cases of FF/SS/FS/SF, the analysis of lithography process variations has to traverse a series of dose and defocus sampling points.

2.1.3.4 Lithography Simulation and OPC Flow

In the typical lithography simulation (as shown in Fig. 2.2), the simulation area is divided into small simulation windows. In practical approach, the window size would be micrometer scale while the chip size would be multi millimeter square. This means the computation with optical model and photo-resist model would be repeated for millions of times over all mask simulation windows, while the process to generate kernels under given dose and defocus variation would only calculated for one time. The computation with optical model (marked as red in Fig. 2.2) would be runtime critical as photo-resist computation is typically of $O(n)$ complexity and the complexity of intensity computation would be from $O(nlog(n))$ to $O(n^2)$ no matter pixel based or pattern edge based algorithms are adopted. n which typically indicates the pixels in a simulation windows could also be of million scale in modern high

quality lithography simulation.

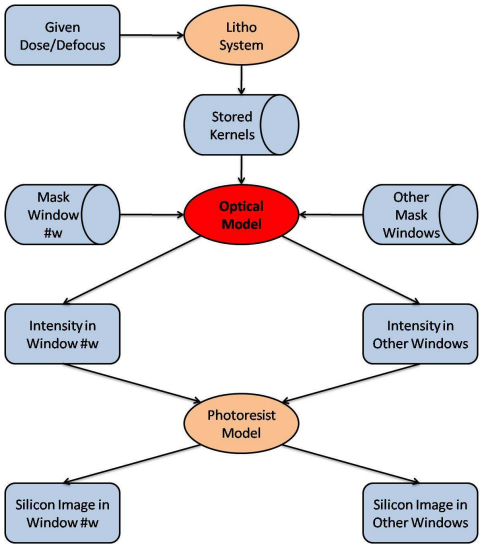


Figure 2.2: Typical Lithography Simulation Flow for Given Process Variation

In the OPC process [37], the mask patter in each window is simulated and modified under normal condition and then variations of different dose and defocus sampling points are simulated for process variation band (PV band) on silicon image (as shown in Fig. 2.3). In this mask generation process, as there are millions of computation windows, lithography simulation and OPC engine are runtime critical steps and also marked in red in Fig. 2.3.

2.1.4 Contribution of the Proposed ILO

In this Chapter, a series of inverse lithography methods will be proposed to better connect design layout to manufacture silicon (L2S) with the context of practical industry standard. The state-of-art $32nm$ lithography system is

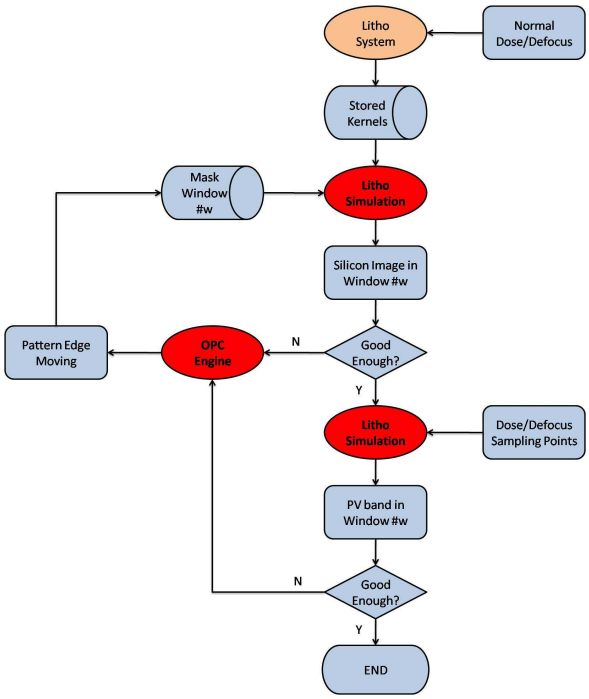


Figure 2.3: Typical OPC Flow for one mask window.

investigated and lithography kernels in Eq. 2.4 are generated by commercial tools with practical process settings where lots of detailed process models (such as focus blurring [37]) are considered with accurate parameter extraction.

Partially coherent optical system is used in the proposed ILO methods as it is the only acceptable optical model for industry application. Binary mask is used as it is valid not only for traditional chrome on glass (COG) mask but also for some phase shift mask (PSM) after modifying the corresponding TCC and kernels (such as the $32nm$ industry lithography system investigated). Constant threshold resist (CTR) model is applied in this chapter and it could be extended to VTR models if that is necessary in the optimization process.

Finally, sum of pixel difference (SPD) is adopted to evaluate the quality of inverse lithography optimization methods. Though the edge placement error (EPE) is commonly used as the metric for OPC process, SPD in fact is equivalent to the total EPE as shown in Fig. 2.4 and Eq. 2.10.

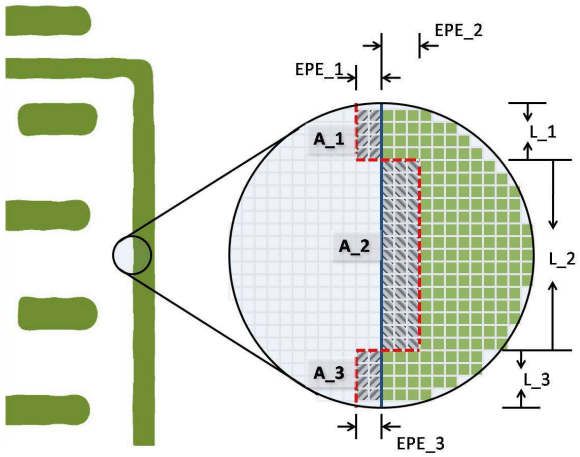


Figure 2.4: The total EPE is equivalent to the sum of pixel difference (SPD).

$$\begin{aligned} EPE_{total} &= \sum EPE_i \times l_i \\ &= EPE_1 \times l_1 + EPE_2 \times l_2 + EPE_3 \times l_3 \\ &= A_1 + A_2 + A_3 \\ &= \sum A_i \\ &= SPD \end{aligned} \tag{2.10}$$

In my proposed methods, two transformations are used to convert the integer optimization problem to a continuo us optimization problem. Compared to earlier attempts in gradient-based inverse lithography [26], significant contributions have been made in the following aspects:

- I use intuitive transformations in the mask transformation, which enables our method to gradually converge to the optimal solution;
- The proposed methods are derived based on partially coherent optical models. Calculating gradient for partially coherent optical models is non-trivial. I developed a highly efficient method for gradient calculation;
- Dose variation is considered into the ILO process with the min-max optimization method and the computation overhead on dose process variation could be omitted.
- The methods are implemented in state-of-the-art industrial lithography environment. The results show that they are feasible in the industrial settings.

The proposed inverse lithography methods recasts the existing heuristics-laden RET methodology into a consistent theoretical framework and produces manufacturable masks. Such a framework could provide means to maximize the full potential of the well-proven optical lithography tools. This is also crucial for future resolution enhancement techniques such as double patterning, because they are built on the basis of lithography-based patterning techniques.

2.2 Efficient ILO under the Normal Condition

Light source of the same $193nm$ wavelength has been used to drive the semiconductor industry from $130nm$ to $22nm$ with six generations. As NA in Eq. 2.1 is hard to increase, different RETs such as OPC, ILT and SMO becomes more critical and promising to reduce k_1 and improve resolution. As runtime is critical for ILO application to practical lithography system, I developed an efficient inverse lithography optimization method under normal lithography condition which is called CORE (computational optimization resolution enhancement) [38]. My ILO method not only has good compatibility with practical industry lithography system but also shows high computational efficiency and stability.

2.2.1 Cost Function

The problem of inverse lithography can be described as to find a mask pattern (among a large number of solutions) so that the final silicon image is as close to a predefined pattern as possible. As illuminated in Fig. 2.4, the sum of pixel difference is equivalent to the total EPE which is used to describe the similarity between the favorite pattern and the silicon image of OPC outputs. For the binary mask or the phase shift mask that could be equivalent to binary mask, the value at each mask pixel can be either 0 or 1. On the other hand, the silicon image is also sampled at certain grid, which can be also regarded

as pixel with value 0 or 1. The sum of pixel difference would be:

$$SPD(\mathbf{M}) = \sum_{j=1}^{MN} |z_j - \hat{z}_j| \tag{2.11}$$

where, $\hat{z}_j$ is predefined as the desired silicon image on pixel j and z_j is the final silicon image on that pixel. M and N represent the number of mask pixels on X and Y dimensions. Vector $\mathbf{M} = (m_1, m_2, \cdots, m_{M\times N})$ represents the mask pattern while m_i is the mask pixel with index i.

In Eq. 2.11, $\hat{z}_j$ and z_j can only be 0 or 1. For the binary mask setting, m_i can also only be 0 or 1. For given lithography system setting and kernels, size of pixels in one window is fixed and M and N would be constant. For most cases and also the adopted $32nm$ industry lithography system, $M = N$.

Because it is hard to calculate gradient of the absolute operator $|\cdot|$ and $\hat{z}_j$ as well as z_j can only be 0 or 1, operator $|\cdot|$ is converted to square operator $(\cdot)^2$ in the cost function and the updated cost function in Eq. 2.12 always has the same value of SPD. This updated cost function is also used in [39].

$$f(\mathbf{M}) = \sum_{j=1}^{MN} (z_j - \hat{z}_j)^2 \tag{2.12}$$

As in $\mathbf{M} = (m_1, m_2, \cdots, m_{M\times N})$ m_i could only be 0 or 1, and $M \times N$ could be of million scale in modern sub-wavelength industry lithography system, the optimization problem of Eq. 2.12 would be integer optimization problem for millions of integer variables. With search space size over $2^{1,000,000}$, any brute-force method will quickly become intractable.

2.2.2 Transformations from $\mathbb{Z}$ to $\mathbb{R}$

After converting SPD in Eq. 2.11 to cost function in Eq. 2.12, transform the original problem to a continuous optimization problem will be meaningful, especially the optical system is inherently continuous and the light intensity onto the photo-resist is also continuous. However, the photo resist model in Eq. 2.9 is discrete and the continuous property is lost from intensity to silicon image. Here based on constant threshold model of photo resist (Eq. 2.9), a Heaviside function is adopted as follow:

$$z_j = S_1(I_j; \alpha_1, t_r) = \frac{1}{1 + e^{-\alpha_1 I_j + \alpha_1 t_r}} \tag{2.13}$$

where I_j is the intensity projected onto the photo resist at pixel j and z_j is the silicon image on wafer on the same pixel with index j. Coefficient t_r works as the photo-resist threshold. The coefficient α_1 defines the slope of the photo-resist model. It has been demonstrated in [36] that this constant threshold photo-resist model can predict the results accurately. The parameter α_1 determines the "sharpness" of the function. The bigger value of α_1, the closer it is to the step function. Furthermore, the Heaviside function $S_1(\cdot)$ is always a C_1 function (with continuous first-order derivatives) which is favorite in many optimization methods.

To realize an efficient inverse lithography optimization, besides transforming the discrete CTR model to the continuous Heaviside function, millions of integer variables as mask pixels is still too excessive to do integer optimization. As the pixels on mask are discrete and there are only two available values

0 and 1 for binary mask (or alternative mask), it is reasonable to convert the original discrete mask $\mathbf{M}$ to a fictitious mask $\mathbf{\Theta}$ which is not only continuous but also in the whole $\mathbb{R}$ space. Furthermore, as the size of $\mathbf{\Theta}$ is the same to $\mathbf{M}$, one-to-one mapping, C_1 function and adjustable approximation would be favorite properties. A cyclic function $m_i = (1 + \cos(\theta))/2$ is used in [26] to transform discrete mask to continuous. However, it is not one-to-one and the "sharpness" is not tunable. In my ILO method, another Heaviside function is used:

$$m_i = S_2(\theta_i; \alpha_2) = \frac{1}{1 + e^{-\alpha_2 \theta_i}} \tag{2.14}$$

where m_i is the discrete mask with value 0 or 1 on pixel i and θ_i is the unconstrained variable of $(-\infty, +\infty)$ for transformation.

With $S_2(\theta_i; \alpha_2)$, original $\{0, 1\}$ mask in integer space will be converted to $(0, 1)$ and then transformed to the whole real space $\mathbb{R}$. From $\{0, 1\}$ to $(0, 1)$, we obtain continuity; from $(0, 1)$ to $(-\infty, +\infty)$, constrained optimization becomes unconstrained. Moreover, one-to-one mapping, C_1 continuity and adjustable "sharpness" are also available. Fig. 2.5 shows how to tune "sharpness" with parameter α_2 in Eq. 2.14.

As discussed in 2.1.3.1, based on Hopkins Equation [32] and partially coherent system, the relationship between intensity and the fictitious mask

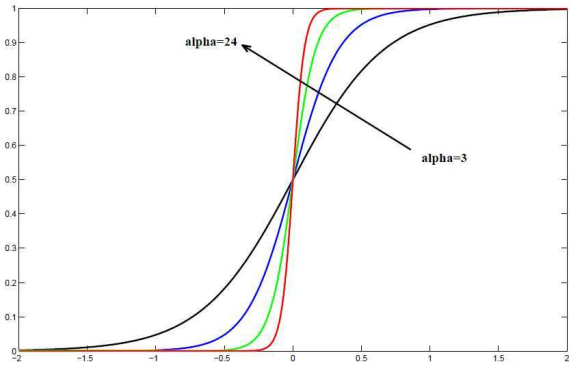


Figure 2.5: The Heaviside function to transform original discrete mask to continuous values with whole $\mathbb{R}$.

can be derived from Eq. 2.6 and Eq. 2.14 as follows,

$$\begin{aligned} I_j = I_j(\mathbf{M}) &= \sum_{k=1}^{n_K} \sigma_k \left| \sum_{i=1}^{MN} h_{ijk} m_i \right|^2 \\ &= I_j(\mathbf{\Theta}) = \sum_{k=1}^{n_K} \sigma_k \left| \sum_{i=1}^{MN} h_{ijk} S_2(\theta_i; \alpha_2) \right|^2 \end{aligned} \tag{2.15}$$

where I_j is the intensity on pixel j, and it is decided by all the mask pixels around it indexed by i, and n_K is the number of kernels to be considered in the partially coherent system.

After applying the Heaviside functions in Eq. 2.5 to the cost function in Eq. 2.12, a new cost function to quantify the goodness of the mask would

be:

$$
\begin{aligned}
f = f(\mathbf{M}) &= \sum_{j=1}^{MN} (z_j - \hat{z}_j)^2 \\
&= f(\mathbf{\Theta}) = \sum_{j=1}^{MN} (S_1(I_j; \alpha_1, t_r) - \hat{z}_j)^2 \\
&= \sum_{j=1}^{MN} \left[S_1 \left(\sum_{k=1}^{n_K} \sigma_k \left| \sum_{i=1}^{MN} h_{ijk} S_2(\theta_i; \alpha_2) \right|^2 ; \alpha_1, t_r \right) - \hat{z}_j \right]^2
\end{aligned}
\tag{2.16}
$$

where $\mathbf{\Theta} = [\theta_1, \theta_2, ...\theta_{MN}]$ defines the fictitious mask with $M \times N$ pixels.

2.2.3 Steepest Descent Method for Partially Coherent Lithography

In order to find $\mathbf{\Theta}$ solution for the optimal object function f, the steepest descent method is applied to continuously modify $\mathbf{\Theta}$ alone the direction pointed by the gradient. Therefore, an efficient method for gradient computation is critical for the steepest decent method. In this part, an efficient calculation for gradient would be first developed for partially coherent lithography system.

Based on Eq. 2.15, we have:

$$
I_j = \sum_{k=1}^{n_K} \sigma_k \left[\sum_{i=1}^{MN} h_{ijk} S_2(\theta_i) \right] \left[\sum_{i=1}^{MN} h^*_{ijk} S_2(\theta_i) \right] \tag{2.17}
$$

where operator $*$ donates complex conjugate, and for any complex number c, $|c|^2 = c \cdot c^*$.

$$\frac{\partial}{\partial\theta_p}f(\Theta) = -2\sum_{j=1}^{MN}(\hat{z}_j - z_j)S_1^{'}(I_j)\sum_{k=1}^{n_K}\sigma_k\cdot \left[\begin{array}{l} \frac{\partial}{\partial\theta_p}\sum_{i=1}^{MN}h_{ijk}S_2(\theta_i)\cdot\sum_{i=1}^{MN}h_{ijk}^{*}S_2(\theta_i) \\ +\sum_{i=1}^{MN}h_{ijk}S_2(\theta_i)\cdot\frac{\partial}{\partial\theta_p}\sum_{i=1}^{MN}h_{ijk}^{*}S_2(\theta_i) \end{array}\right] \tag{2.18}$$

For any $p \neq i$, $\frac{\partial}{\partial\theta_p}S_2\left(\theta_i\right) = 0$, and we have

$$\frac{\partial}{\partial\theta_p}\sum_{i=1}^{MN}h_{ijk}S_2(\theta_i) = h_{pjk}S_2^{'}(\theta_p) \tag{2.19}$$

Similarly:

$$\frac{\partial}{\partial\theta_p}\sum_{i=1}^{MN}h_{ijk}^{*}S_2(\theta_i) = h_{pjk}^{*}S_2^{'}(\theta_p) \tag{2.20}$$

Combine Eq. 2.19 and 2.20 with Eq. 2.18, we have

$$\begin{aligned} \frac{\partial}{\partial\theta_p}f(\Theta) =& -2\sum_{j=1}^{MN}(\hat{z}_j - z_j)S_1^{'}(I_j)\sum_{k=1}^{n_K}\sigma_k\cdot \\ & \cdot\left[h_{pjk}S_2^{'}(\theta_p)\sum_{i=1}^{MN}h_{ijk}^{*}S_2(\theta_i) + \sum_{i=1}^{MN}h_{ijk}S_2(\theta_i)h_{pjk}^{*}S_2^{'}(\theta_p)\right] \\ =& -2S_2^{'}(\theta_p)\sum_{k=1}^{n_K}\sigma_k\sum_{j=1}^{MN}(\hat{z}_j - z_j)S_1^{'}(I_j)\cdot \\ & \cdot\left[h_{pjk}\sum_{i=1}^{MN}h_{ijk}^{*}S_2(\theta_i) + \sum_{i=1}^{MN}h_{ijk}S_2(\theta_i)h_{pjk}^{*}\right] \\ =& -2S_2^{'}(\theta_p)\sum_{k=1}^{n_K}\sigma_k\sum_{j=1}^{MN}(\hat{z}_j - z_j)S_1^{'}(I_j)\cdot \\ & \cdot\left[h_{pjk}\sum_{i=1}^{MN}h_{ijk}^{*}S_2(\theta_i) + \sum_{i=1}^{MN}h_{ijk}S_2(\theta_i)h_{pjk}^{*}\right] \end{aligned} \tag{2.21}$$

Note that the most inner summation actually defines the field at location j due to kernel k. Then the field at location j is E_{jk}:

$$E_{jk} = \sum_{i=1}^{MN} h_{ijk} m_i = \sum_{i=1}^{MN} h_{ijk} S_2(\theta_i) \tag{2.22}$$

$$\mathbf{E}_k = E_{*k} = h_{**k} \otimes m_* = h_{**k} \otimes S_2(\theta_*) \tag{2.23}$$

where E_{jk} is the field of kernel k on pixel j and $\mathbf{E}_k$ is the field matrix from kernel k. Combining Eq. 2.22 with Eq. 2.21, we have:

$$\frac{\partial}{\partial \theta_p} f(\Theta) = -2 \sum_{k=1}^{n_K} \sigma_k S_2'(\theta_p) \sum_{j=1}^{MN} (\hat{z}_j - z_j) S_1'(I_j) \cdot \left(h_{pjk} E_{jk}^* + E_{jk} h_{pjk}^* \right) \tag{2.24}$$

We can define term C_{pk} for pixel p under kernel k as:

$$C_{pk} = \sum_{j}^{MN} (\hat{z}_j - z_j) S_1'(I_j) \cdot h_{pjl}^* E_{jk} \tag{2.25}$$

And notice,

$$h_{pj} E_{jk}^* + h_{pj}^* E_{jk} = 2Real\{h_{pj} E_{jk}^*\} = 2Real\{h_{pj}^* E_{jk}\} \tag{2.26}$$

Where $Real\{\cdot\}$ donates the real part of a complex number. The term C_{pk} can be calculated very similar to the calculation of field E_{jk} from kernel k though we have to conjugate the kernel h_{pjl} or the filed E_{jk}. we can follow the same procedure of convolution and calculate all of them in one operation, rather than calculating C_{pk} one by one. Afterwards, we can calculate the Jacobian as:

$$\frac{\partial}{\partial \theta_p} f(\Theta) = -4 \sum_{k=1}^{n_K} \sigma_k S_2'(\theta_p) \cdot Real\{C_{pk}\} \tag{2.27}$$

After replacing C_{pk} in Eq. 2.27 with Eq. 2.25, the gradient on each pixel is:

$$\frac{\partial f}{\partial \theta_p} = 4 \cdot \sum_{k=1}^{n_K} \sigma_k \cdot S_2' \cdot Real \left\{ \sum_{j=1}^{MN} [S_1'(z_j - \hat{z}_j) (E_{jk})^*] \cdot h_{pjk} \right\} \tag{2.28}$$

where $S_1' = S_1'(I_j)$, $S_2' = S_2'(\theta_p)$, and $E_{jk} = \sum_{i=1}^{MN} (h_{ijk} m_i)$ can be reused as it is calculated and stored during of silicon image computation. Be aware that S_1' is indexed by j and must be kept inside of $\sum_{j=1}^{MN} [(z_j - \hat{z}_j) S_1' (E_{jk})^*] \cdot h_{pjk}$, while S_2' is just a function of θ_p which is only indexed by p and so S_2' could be taken out of the summary.

And the gradient of all pixels (the Jacobian matrix) is shown as follows:

$$\nabla f = 4 \cdot \sum_{k=1}^{n_K} \sigma_k \cdot S_2' \odot Real \left\{ \left[S_1' \odot (z_j - \hat{z}_j) \odot E_{jk}^* \right] \otimes h_{*jk} \right\} \tag{2.29}$$

where operator $\otimes$ is the inner product and it computation complexity for matrices is $O(n)$.

Hereby, we can see that the computation of the gradient matrix ∇f could be very efficient as the key calculation is finished with only one more convolution which has the same complexity for the forward simulation of silicon image.

After getting the gradients, there are ways to do decant searching and update $\mathbf{\Theta}$ from step (ω) to $(\omega + 1)$. One of the most straightforward ways is the steepest decent method as:

$$\mathbf{\Theta}_{\omega+1} = \mathbf{\Theta}_{\omega} - \alpha_3 \cdot \nabla f_{\omega} \tag{2.30}$$

where α_3 is some small positive number specified by the users.

2.2.4 Complexity

Even though the two transformations succeed to convert integer optimization into unconstrained continuous optimization, without efficient calculation of gradient for the practical partially coherent lithography system, the inverse lithography optimization is still not practical for state-of-art industry application. Eq. 2.18 shows the most straightforward gradient for each pixel. Generally, to calculate the intensity on one pixel, the computation complexity is $O(n)$; for the intensity of all pixels, the complexity is $O(n^2)$ which is also the complexity to calculate the cost function as photo-resist model is very local; the gradient of cost function over one pixel can be $O(n^2)$ but the Jacobian for all pixels would be $O(n^3)$. For a computational window with $M \times N$ pixels, the complexity to calculate gradients of all pixel would be as much as $O\left((MN)^3\right)$ with $O(n^3)$ complexity. As MN would be of million scale, after converting integer optimization into unconstrained continuous optimization, complexity is reduced from scale of $2^{1,000,000}$ to $1,000,000^3$, but still not affordable for industry application.

Eq. 2.19 and 2.20 help to avoid one sum of all pixels and the gradient of one pixel in Eq. 2.21 is still of $O(n^2)$ complexity while the computation is more effient. With the reuse the computation result of field matrix in Eq. 2.23, Eq. 2.28 would have $O(n)$ complexity for each pixel and Eq. 2.29 would have $O(n^2)$ complexity at most for all pixels. As the computation of the field matrix in Eq. 2.23 has also $O(n^2)$ complexity at most. Therefore, The $O(n^3)$ complexity problem is successfully decomposed to two independent problems

Table 2.2: The complexity to calculate Jacobian.

Methods	Complexity
Original	$O(n^3)$
Mask transformation	$O(n^3)$, more efficient
Reuse light field	$O(n^2)$
FFT	$O(n \log n)$
$h_{pj}E^*_{jk} + h^*_{pj}E_{jk} = 2Real\left\{h_{pj}E^*_{jk}\right\}$	$O(n \log n)$, more efficient

with $O(n^2)$ complexity.

Furthermore both of the two $O(n^2)$ complexity problems are based on matrix convolution which can be reduced to $O(n \log n)$ after using FFT to replace direct convolution. Eq. 2.29 also has a few inner production operators $\odot$ which are not critical as the complexity is $O(n)$. So the overall complexity at each step of computing gradients is $O(MN \log(MN))$. Many contemporary FFT packages are highly efficient. For example, using FFTW [40], the FFT of 2M floating point variables only takes about 60 milliseconds on a typical x86 architecture computer, while a gradient calculation takes merely seconds to complete.

2.3 Dose Process Variation Aware ILO

The aggressive scaling down has pushed the manufacturing process to combined use of high NA as well as low k_1 factor, and induced significant shrinking of tolerable lithography process window. In quite a few research and EDA tools of OPC and lithography verification, the process variations have

been considered [37, 41–44]. However, there are very limited publications on process aware inverse lithography [27].

Dose process variation aware inverse lithography optimization (ILO) would be discussed in this section. To keep the efficiency of computation and optimization, we used continuous mask ILO, constant threshold resist (CTR) model, linearity approximation of intensity-dose relation, first order approximation between cost function and dose variations, and min-max optimization techniques. The optimization target is to make the final mask generate silicon image as close as possible to layout even at the specific dose-defocus variation corners (worst cases).

2.3.1 Model of Dose Variations

Different from the depth of focus variations, the basic physical model of dose variation is simple and straightforward if we ignore the aberrations caused by lens heating. Moreover, if we regard dose variation as the first order approximate impact of different kinds of process variations onto CD as [45], the linear model of effective dose variation shown as below will be reasonable.

$$\mathbf{I} = \mathbf{I_0} \cdot (1 + \delta_d) \tag{2.31}$$

where $\delta_d = \Delta d/d_0$ is the relative dose variations. For each pixel, it is,

$$I_j = I_{j,0} \cdot (1 + \delta_d) \tag{2.32}$$

- There the variation of dose is constant over the whole simulation window, meaning there is no within-window dose variation.

- Constant threshold photo-resist (CTR) model is used.

The assumptions above are also shown in Fig. 2.6. The intensity is a continuous function at location x, and its intersection with CTR_0 is the location of pattern edge and decides critical dimension (CD). The original light intensity is I_0, constant photo-resist threshold is CTR_0 and the corresponding pattern edge is x_0. If there is dose variations, the light intensity might change to I and the edge will move to x_1. A new CTR with the same intensity I_0 will move the edge from x_0 to x_2. By tuning CTR, x_2 can overlap with x_1. Therefor to get the wafer image under dose process variations, modifying and calibrating current intensity I or photo-resist threshold CTR are equivalent.

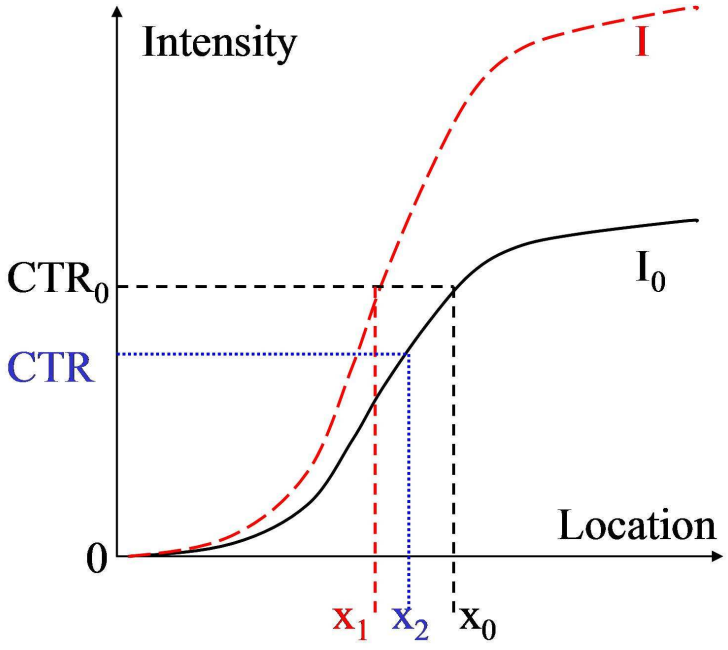


Figure 2.6: CD variations caused by dose uncertainty.

2.3.1.1 Recalculate Wafer Image with Dose Variation

As shown in Fig. 2.6, the dose variations will cause the shift of intensity (from the black solid line to the red dash line) and also movement of pattern edge (from x_0 to x_1). The original edge is at x_0 where intensity I is equal to constant threshold resist CTR_0, where $I_0(x_0) = CTR_0$. When there is dose variation, the new location of edge x_1 is calculated by

$$I(x_1) = CTR_0 \tag{2.33}$$

where

$$I(x_1) = (1 + \delta_d) \cdot I_0(x_1)$$

This is equivalent to:

$$I_0(x_1) = \frac{CTR_0}{1 + \delta_d} \tag{2.34}$$

Instead of shifting intensity, we can also to move the photo-resist threshold from CTR_0 to a new value while keeping the intensity to be the same. If the new location x_2 is decided by:

$$I_0(x_2) = CTR \tag{2.35}$$

and if we set

$$CTR = \frac{1}{1 + \delta_d} CTR_0 \tag{2.36}$$

x_2 calculated from Eq. 2.35 will be equal to x_1 from Eq. 2.34.

Therefore in order to calculate the wafer image with dose variations, there is no difference to modify intensity or CTR. However, in the rest of

this section, I will show that the two methods might be slightly different in calculating gradients for min-max problem optimization.

2.3.1.2 Introduction of Min-max Optimization

Besides statistical optimization [46, 47], Min-max optimization [48, 49] is also used widely to circuit optimization [50]. The min-max optimization problem in inverse lithography optimization with dose and defocus process variations can be expressed in Eq. 2.37.

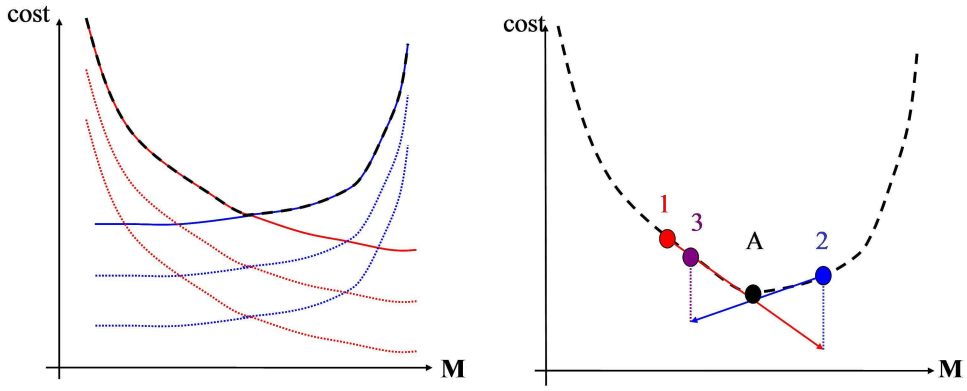
$$\min_{\mathrm{M}} \max \left\{ f\left(M; \delta_{\mathbf{d}}, \delta_{\mathbf{f}}\right) \right\} = \min_{\mathbf{M}} \left\{ F_{\max}\left(M; \delta_{\mathbf{d}}, \delta_{\mathbf{f}}\right) \right\} \tag{2.37}$$

where

$$F_{\max}\left(M; \delta_{\mathbf{d}}, \delta_{\mathbf{f}}\right) = \max_{\delta_d, \delta_f} \left\{ f\left(M; \delta_d, \delta_f \,|\delta_d = \delta_{d,i}, \delta_f = \delta_{f,i}\right) \right\}$$

The min-max optimization for the process variation window can be divided into two steps: 1. calculate maximum cost function F_{max} through all process corners; 2. do mask optimization based on F_{max}. As shown in Fig. 2.7(a), among the blue and red lines under different dose variations, only the solid blue line and the solid red line do matter to this min-max optimization as they compose the F_{max} curve shown in black dash line. The intersection of the blue and red solid lines is the singular point and when the mask optimization reach around the singular point, the converging process might be very slow as the oscillating trend shown in Fig. 2.7(b) [49]. In [49], the min-max optimization is divided into two steps: 1. first-order optimization when the solution is far away from singular [48]; 2. approximate second-order for singu-

lar solution. There are several times of switching between above two methods during the min-max optimization [49].



(a) Cost function curves at different dose variations.

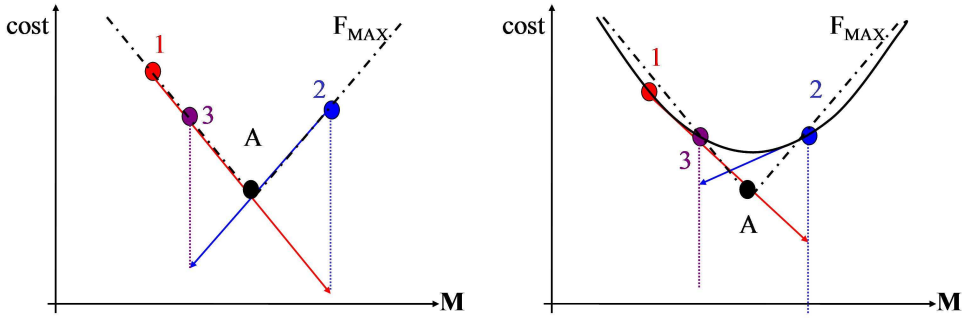
(b) Oscillation around the singular point.

Figure 2.7: Min-max optimization for different dose variations.
Because it is more critical to find the worst cases over the whole dose variation window rather than all the red and blue curves of cost function, only the contour with the dash black line needs to be optimized. Vector **M** is to indicate the combination of all mask pixels.

Even though the above min-max optimization technique has been successful used for some microwave circuit optimization [50], there is very limited work on the min-max problem for process variation window based ILO. In process window aware ILO, variables δ_d and δ_f in max part (F_{max}) are independent to the continuous mask optimization in min part. While the dose process variation models are mature (Eq. 2.31), there are no accurate and reliable models for defocus variations. Currently, TCC and kernel sets for each

depth-of-focus have to been modeled separately. [42] models the defocus process variation's impact on silicon image as quadratic functions, however there is not way to guarantee the model error boundary. Thus in the following parts, the min-max optimization will focus on dose variation aware ILO for mature and reliable dose variation model is available.

As shown in Fig. 2.8(a), $F_{\max}$ is made up of two functions $f_{dose=d\max}$ and $f_{dose=-d\max}$. At the intersection, it is the singular point [50], and as shown in Fig. 2.8(a), around singular point the converging process would be much slower for its oscillating gradients.



(a) Sharp transform of maximum cost function curve around a singular point.

(b) Maximum cost function curve is smoothed by S_3' function.

Figure 2.8: Gradient searching iterations around a singular point.

One potential solution to deal with the converging issue is to do some smoothing around the singular point. As shown in Fig. 2.8(b), when the gradient searching is close to the singular point, based on the smoothed cost

function, the recalculated gradients would be smaller and the contribution of ζ_d in 2.47 to the total gradients would be depressed, and the oscillation around singular point would be reduced.

2.3.2 Dose Variation Window Aware ILT

In this part, dose variation aware inverse lithography optimization will be presented. In forward lithography simulation with the consideration of dose variations, there is no wafer image difference to modify lithography intensity or the photo resist threshold in CTR. However, in the inverse lithography min-max optimization. The gradients from modifying CTR and intensity will be different. The gradient for ILO with dose variations based on CTR will be first presented, then the gradient based on intensity will be deducted, between which certain difference could be identified. Whereas the choice of the right gradient for min-max optimization is dependent on the parameter extraction and calibration of dose variation. Fundamentally, dose variation is an equivalent expression of first order (linear) effect of lithography process variations. In the following experiment part, ILO gradient with dose variations would be based on the intensity change caused by dose variations.

2.3.2.1 Using Equivalent CTR in Min-max Optimization for Dose Variation Window

As shown in 2.35, the wafer image difference caused by dose variations can be also calculated by new threshold of variations in 2.36.

$$t_r\left(\delta_d\right) = \frac{1}{1+\delta_d} t_{r0} \tag{2.38}$$

Based on the above dose variations induced threshold model, the cost function under dose variations in 2.44 would become:

$$f \simeq f_0 + 2\delta_d \cdot \sum_{j=1}^{MN} S_1' \cdot \left(z_{j,0} - \hat{z}_j\right) \cdot t_{r0} \cdot \left(1+\delta_d\right)^{-2} \tag{2.39}$$

and 2.46 will be,

$$F_{\max}\left(M; \delta_{\mathbf{d}}, 0\right) \approx f_0 + \frac{2t_{r0} \cdot \delta_{d,\max}}{\left(1-\delta_{d,\max}\right)^2} \cdot \left| \sum_{j=1}^{MN} S_1' \cdot \left(z_{j,0} - \hat{z}_j\right) \right| \tag{2.40}$$

where t_{r0} is the threshold of resist at nominal condition without dose variation.

And 2.47 will be:

$$\begin{aligned} &\nabla F_{\max}\left(M; \delta_{\mathbf{d}}, 0\right) \\ &= \sum_{k=1}^{n} \sigma_k \cdot 4 \cdot S_2' \odot Real \left\{ \left[\begin{array}{c} S_1' \odot \left(z_j - \hat{z}_j\right) \\ + \dfrac{\delta_{d,\max}}{\left(1-\delta_{d,\max}\right)^2} \cdot \zeta_t \end{array} \right] \odot E_{jk}^{*} \otimes h_{pjk} \right\} \end{aligned} \tag{2.41}$$

where ζ_t is different from ζ_d in 2.48:

$$\zeta_t = t_{r0} \cdot S_3' \cdot \left(S_1'' \odot \left(z_{j,0} - \hat{z}_j\right) + S_1' \odot S_1'\right) \tag{2.42}$$

2.3.2.2 Estimation Maximum Cost Function with Dose Variations

After putting dose and defocus variations into consideration, the cost function in Eq. 2.16 at certain process corner is changed to,

$$f = f\left(\mathbf{M}; \delta_d, \delta_f\right) = \sum_{j=1}^{MN} \left(z_j\left(\mathbf{M}; \delta_d, \delta_f\right) - \hat{z}_j\right)^2 \tag{2.43}$$

If we only consider the dose variations, based on Eq. 2.32 and first order approximation, we have

$$f = f_0 + 2\delta_d \cdot \sum_{j=1}^{MN} S_1'\left(z_j - \hat{z}_j\right) I_j\left(\mathbf{M}; 0, 0\right) \tag{2.44}$$

where

$$f_0 = f\left(\mathbf{M}; 0, 0\right) = \sum_{j=1}^{MN} \left(z_j\left(\mathbf{M}; 0, 0\right) - \hat{z}_j\right)^2 \geq 0 \tag{2.45}$$

.

Thus, $F_{\max}$ is,

$$F_{\max}\left(M; \delta_{\mathbf{d}}, \delta_{\mathbf{f}}\right) = f_0 + 2\delta_{d,\max} \cdot \left| \sum_{j=1}^{MN} S_1'\left(z_j - \hat{z}_j\right) I_j\left(\mathbf{M}; 0, 0\right) \right| \tag{2.46}$$

where we assume $\delta_d \in [-\delta_{d,\max}, \delta_{d,\max}]$ and at nominal condition, $\delta_d = 0$.

2.3.2.3 Min-max Optimization for Dose Variation Window

Based on $F_{\max}$, the gradient of matrix of $\nabla F_{\max}$ would be calculated to replace ∇f in Eq. 2.29.

$$\begin{aligned} &\nabla F_{\max}\left(M; \delta_{\mathbf{d}}, 0\right) \\ &= \sum_{k=1}^{n} \sigma_k \cdot 4 \cdot S_2' \odot Real\left\{\left[S_1' \odot \left(z_j - \hat{z}_j\right) + \delta_{d,\max} \cdot \zeta_d\right] \odot E_{jk}^* \otimes h_{pjk}\right\} \end{aligned} \tag{2.47}$$

where, ζ_d is introduced for the optimization at dose variation corners. It is,

$$\zeta_d = S_3' \cdot (S_1'' \odot (z_{j,0} - \hat{z}_j) \odot I_{j,0} + S_1' \odot S_1' \odot I_{j,0} + S_1' \odot (z_{j,0} - \hat{z}_j)) \quad (2.48)$$

S_3 is the absolute function.

$$S_3(x) = |x|$$

Its gradient S_3' is,

$$S_3' = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases} \quad (2.49)$$

And smoothing function can also been used to replace absolute function to get continuous gradient. Then S_3 would be,

$$S_3(x; \alpha_4) = \frac{1}{\alpha_4} \ln\left(1 + e^{2\alpha_4 x}\right) - x \cong |x| \quad (2.50)$$

Then its gradient S_3' will be renewed to,

$$S_3' = \frac{d}{dx} S_3(x; \alpha_4) = 3 - \frac{4}{1 + e^{2\alpha_4 x}} \quad (2.51)$$

The detailed mathematic deduction can be found in Appendix 1.3.

The optimization process of inverse lithography are divided into two parts: global and detail ILO. In the global part, ILO is to get reasonable results roughly close to the optimal solution. In the detail part, the solution under nominal condition is close to the optimal (cost function of f_0 would not be too big). The dose process variation window based ILO is finished in the detail part.

2.3.2.4 Complexity of Min-Max Optimization with Dose Variations

As shown in Eq. 2.48, the computation ζ_d is made up of inner multiplication and the complexity is n. As there is only one convolution in Eq. 2.47, the total computational complexity of gradient in min-max optimization could be $n \log n$, the same to Eq. 2.29.

2.4 Experimental Results

In this section, experimental results will be presented including mask optimization in both the normal condition and variable dosage conditions.

2.4.1 Mask Optimization of 32nm M1 Technology

The proposed mask optimization method has been implemented in a state-of-the-art industrial lithography environment at 32nm node. Partially coherent models of the lithography system are used. The optimization procedure is implemented in C, using FFTW [40] as the Fast Fourier Transformation engine. The first example is the lower layer metal routing from an industrial design. Fig. 2.9 shows the intended target.

After the completion of our optimization steps, the final light intensity on the wafer surface is shown in Fig. 2.10, which translates into the final binary wafer image shown in Fig. 2.11.

The final mask is shown in Fig. 2.12. Note that the mask is drastically different from a mask generated by traditional RET techniques. There is a

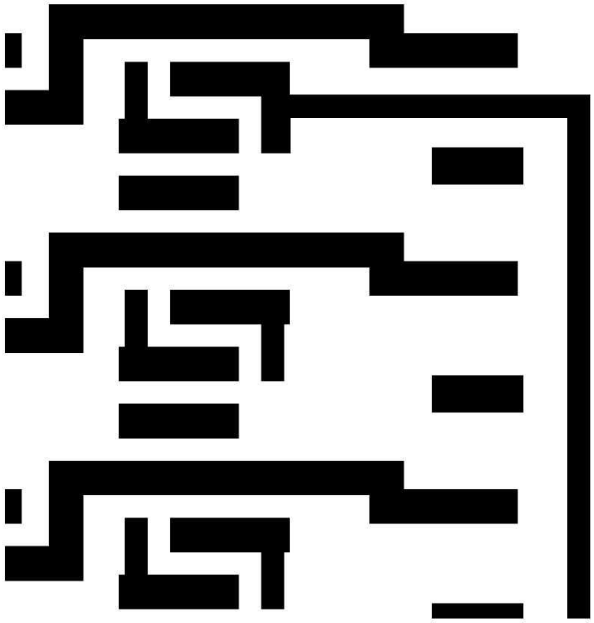


Figure 2.9: Case #1: 32nm layout of metal 1 layer.

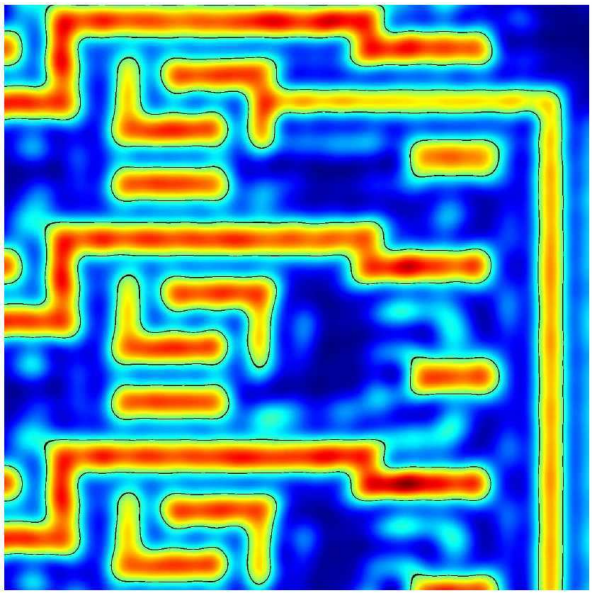


Figure 2.10: Case #1: 32nm intensity distribution of metal 1 layer the proposed ILT.

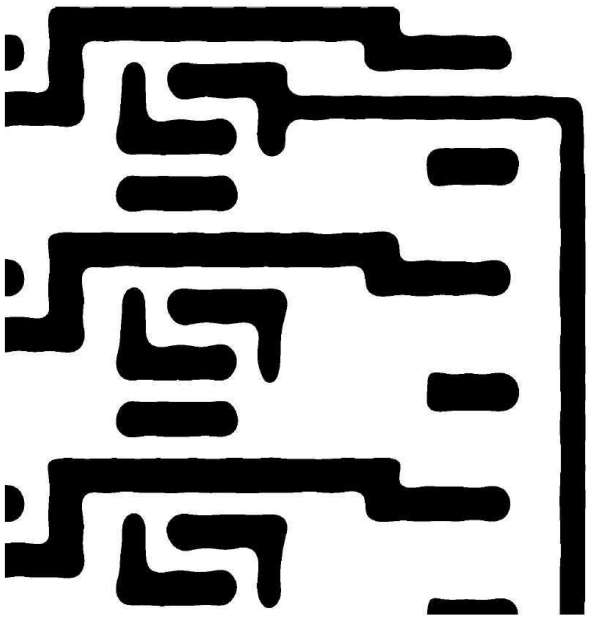


Figure 2.11: Case #1: 32nm silicon image of metal 1 layer after using the proposed ILT.

thin vertical bar on the right hand side of the mask which is to facilitate the printing of the true vertical wiring in the target. To some extent, it is very similar to a SRAF. However, the traditional SRAF is added by applying rules, and the bar in our mask is automatically generated by the optimization.

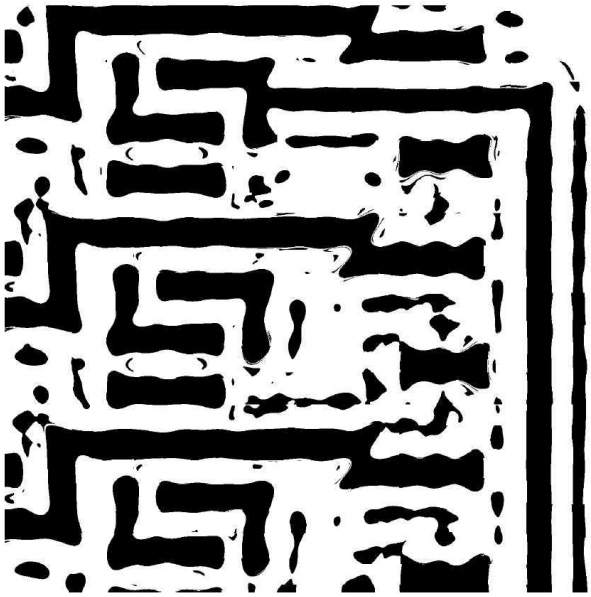


Figure 2.12: Case #1: 32nm mask of metal 1 generated by the proposed ILT. final mask generated by the optimization process. Note many small extra features which function as SRAF but are automatically generated by the optimization routine.

The convergence of the objective function F is shown in Fig. 2.13. For this particular example, it takes approximate 180 iterations to achieve convergence. The first 10% iterations achieves most of the error reduction.

The second example is a denser pattern on the metal layer. The target is shown in Fig. 2.14 and the light intensity of the final mask is shown in

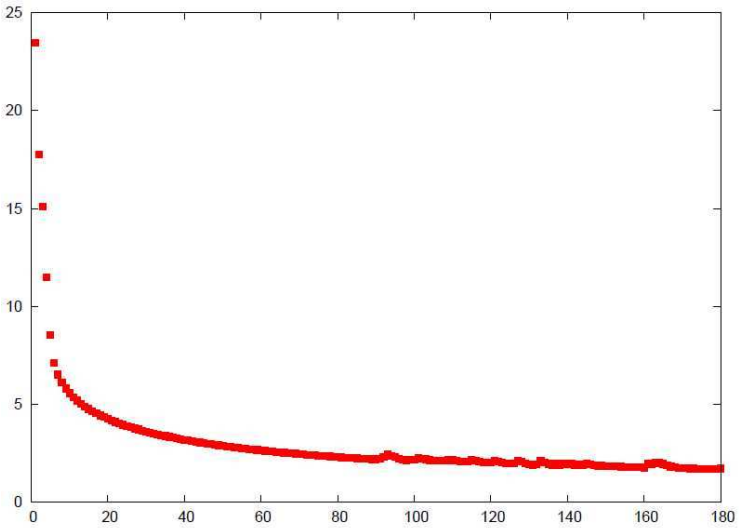


Figure 2.13: Convergence of the algorithm for Case #1.

Fig. 2.15.

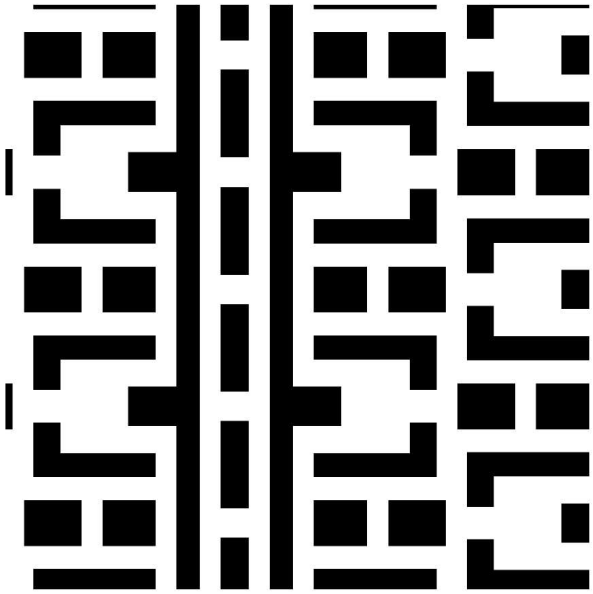


Figure 2.14: Case #2: 32nm layout of metal 1 with denser pattern.

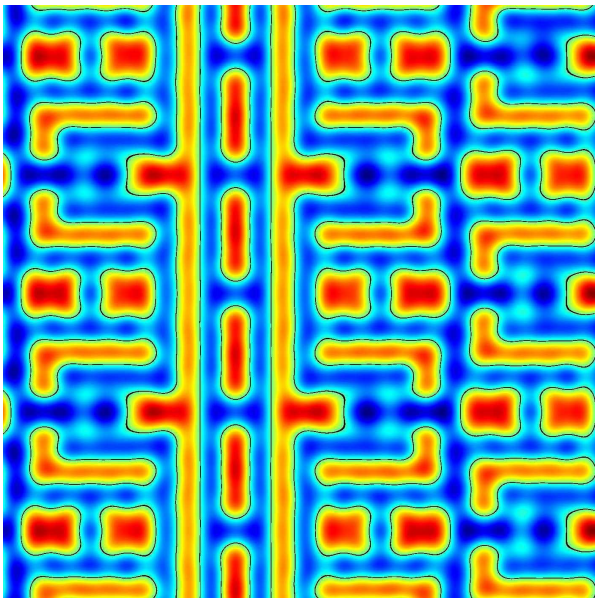


Figure 2.15: Case #2: 32nm intensity distribution of metal 1 from the final mask calculated by the proposed ILT.

2.4.2 Mask Optimization of $22nm$ M1 Layout with $32nm$ Technology

Finally, we demonstrate a "pushed" example. It is artificially generated by scaling pattern shown in the first example so that it mimics the design of the next generation technology. In other words, we are trying to use $32nm$ lithography tool to print a $22nm$ layout design. The target is shown in Fig. 2.16 and the final wafer image is shown in Fig. 2.17.

As shown in Fig. 2.17, the results coincide well with the designed layout. This indicates that even the lithography wavelength remains constant, theoretically inverse lithography can push the technology one generation ahead.

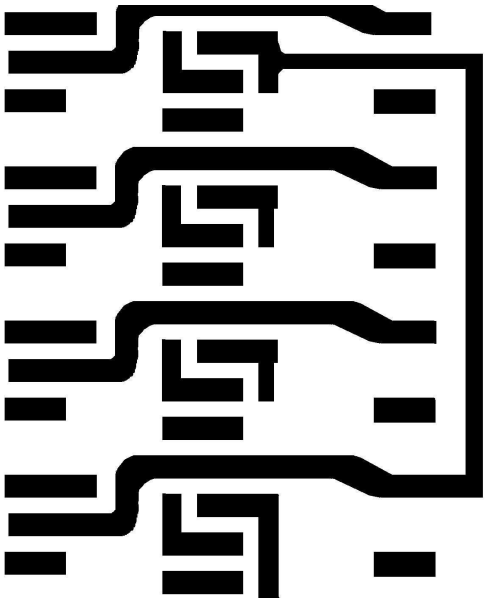


Figure 2.16: Case #3: pushed targeting layout.
Simple scaling of the target shapes from the Generation N to Generation $N+1$.

2.4.3 Dose Variation Window Aware ILT for $32nm$ M1

Based on a given dose variation window, the dose variation aware ILT has been applied. Fig. 2.18 shows the decreasing of cost function. The first part before iteration #520 of ILT is based on normal condition without dose variations. After that, during detail ILO, dose variation window aware ILT is applied. The jump of cost function around iteration #520 after considering dose variations window shows the different of cost function between normal and dose process window condition. After applying dose variation window aware ILT, the cost function of mismatching between layout and silicon image is obviously reduced also at dose variation corners.

The details around iteration #520 during the period that dose variation

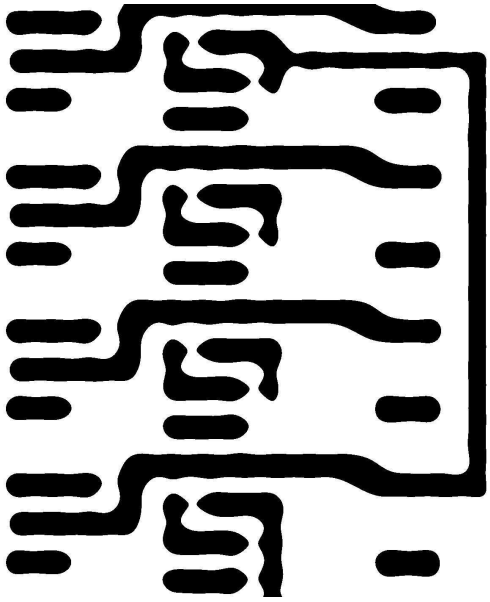


Figure 2.17: Case #3: silicon image from the final mask.
The proposed ILT is applied to extend 32nm lithography technology to 22nm layout design.

window aware ILT is applied are shown in Fig. 2.19. As the cost function is changed and optimization strategy is modified, there are some obvious swing of cost function. These swings shown the difference between normal condition ILT and dose variation window aware ILT.

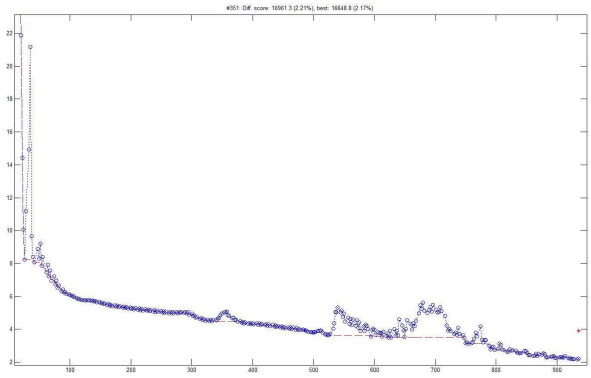


Figure 2.18: Case #4: Cost function decline of dose variation window based ILO.
The first part before iteration #520 is global optimization without dose variation consideration; the second part is detail ILO where process corners of dose variation are considered.

2.5 Summary

In this chapter, we present an efficient inverse lithograph optimization method for deep sub-wavelength lithography. The method utilize two transformations and steepest descent to achieve good convergence speed. We also derived a method to efficiently calculate the gradient using FFT. The experimental results in state-of-the-art industrial lithography environment demonstrate that the method is highly effective.

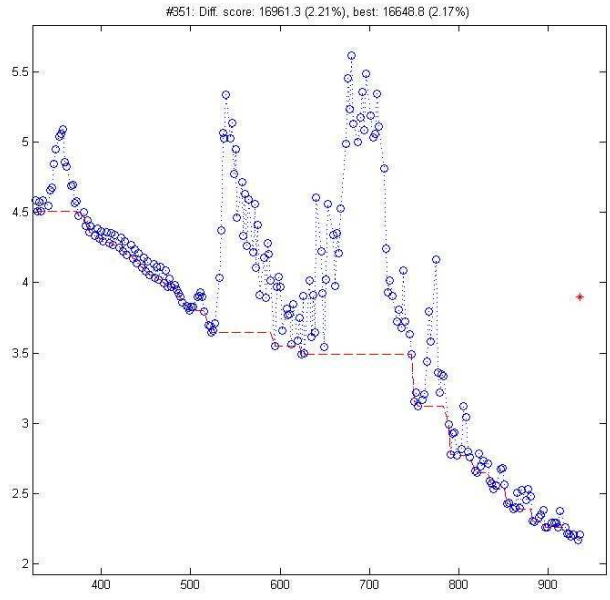


Figure 2.19: Case #4: details of cost function decline of dose variation window based ILO.
The oscillation of cost function becomes dramatic after dose variations are considered in detail ILO.

Chapter 3

Devices with Non-rectangular Gates

3.1 Challenges of Nanoscale Devices

The continuously scaling down guided by Moore law has driven device feature size to nanometer scales with lots of advanced processing technologies. Two kinds of challenges would be remarkable for the nanoscale device modeling and optimization: some traditional non-critical issues of large scale (sub-micrometer) would become significant at nanometer scale; new process technology needs to be modeled and embedded to the optimization strategy. In this chapter, research on both above aspects would be investigated.

Non-rectangularity of device's gates caused by gate roughness and so on. has been investigated for sub-micrometer scale devices. However, as the gate roughness and other non-rectangularity could not scale down with the gate length, their impact on nanoscale device performance is becoming notable. In this chapter, non-rectangular gate would be investigated and a compact model and simulation flow for non-rectangular gates would be first proposed.

The non-rectangularity is not only from gate roughness but also caused by adopting sub-wavelength lithography technology. With scaling down to sub-wavelength technology node such as 65nm, 45nm, 32nm, 22nm and below,

even after advanced lithography technologies, e.g. optical proximity correction (OPC) and double pattern, the gate may still be non-rectangular. There are several limited works on the device and circuit characterizations for the post-OPC non-ideal-shape silicon images. However, most of them target to find the equivalent gate length (EGL). Different EGLs have to be used for timing and leakage, and thus it is hard to to be used for coherent circuit simulations. Thus, I proposed post-litho device model of non-rectangular gates and circuit simulation.

As scaling down on feature size has been more difficult, another advanced techniques to improve device performance are adopted. Strained silicon is one of the most widely used processing techniques in nanoscale IC manufacturing. By stressing the silicon in channels, the mobility of both NMOS's electrons and PMOS's cavities can be significantly improved. When stressing engineer is used for strained silicon, the size of the active area might greatly affect the mobility of cavities and electrons. An optimization strategy of strained silicon PMOS at detail placement step is first proposed in my DATE paper and would be also discussed in this chapter.

3.2 Motivation of Post-Litho Device Modeling for Non-Rectangular Gates

After moving to nanometer scale dimensions with higher densities, IC technology is challenged by the rapidly reducing printability of fine lithographic patterns with rectangular shapes due to the fundamental limitation of

both microlithography systems and process variations. As $193nm$ lithographic system is still used to print critical dimension (CD) of $65nm$ (and likely $45nm$, $32nm$ or even below feature size [51]) with the difficulty of applying electron beam or 157nm lithography into mainstream in near future [13, 52, 53], various resolution enhancement techniques (RET), including optical proximity correction (OPC) [15, 16], phase shift mask (PSM), off-axis illumination (OAI) and double pattern are applied to push the lithographic systems to their limits [52]. Even with extensive RETs, the gate shapes may still be away from perfect rectangles which affect the timing of the whole chip [54]. Meanwhile, different process variations, such as dose, focus, etching variations, mask alignment error, can push the manufactured silicon image (SI) of the gate even further away from the design-intended rectangle layout. Moreover, after photo resist ashing [55–57] (also called oxide lateral etching [58] or resist thinning [59]), constant absolute value of gate length roughness will account for much bigger ratio related to the reduced gate length. In fact, as the channel region is determined by the interfaces between dopant profiles of channel and source/drain under the poly gate, the recent paper [60] has shown that even under rectangular gate, the channel region may be seriously still non-rectangular, which makes the rectangular gate effectively non-rectangular. As the continuous shrinkage of feature sizes, non-rectangular gates and channels are unavoidable due to the limitation of manufacturing process and have to be dealt with and simulated carefully.

The electrical and performance impact of the non-rectangular gate

shape has received a lot of research attention recently. It was first treated as random variations on the edges of the gate, i.e., line edge roughness (LER) [60–64]. The devices with LER can be simulated by 3-D TCAD (process simulation) methods but they are too slow to be performed for circuit level simulations and optimizations. A commonly used technique is gate slicing which applies 2-D TCAD to study LER [62, 63]. As the non-rectangular of channel shape is mostly decided directly by the gate shape, the non-rectangular gate can be directly used to predict the channel shape. Recent works [65–67] use gate slicing and equivalent gate length (EGL) methods to simulate the impact of non-rectangular gate shapes in SPICE, which are much faster than TCAD software. Two EGLs of each device are defined to replace the original uniform gate length set on layout: ON EGL for timing issues when the device is turned on and OFF EGL for leakage issues when the device is cut off.

Although ON/OFF EGLs can well model a non-rectangular device in either of its two specific working states, they are hard to be used for coherent circuit simulations in practice, since it is often difficult to tell when and which devices are absolutely on or off for complicated cell schematics, thus it is hard to choose the right EGL for circuit simulation. Another limitation of these EGL works [65–67] is that they only studied the performance difference between non-rectangular and ideal rectangular gates. In fact, device models, such as BSIM3 [68], BSIM4 [68], and PTM [69], and their parameters are based on and extracted from the real devices which are made up of certain non-rectangular shapes. So the impact of gate shape with certain non-rectangles

on device performance has been already included in the device modeling process [67]. Performance simulations of gates between different non-rectangular shapes are necessary and crucial, otherwise the impact of non-rectangular gate shape may be overestimated.

As an essential bridge between manufacturing process and the circuit design, all of the widely used compact device models [70] nowadays such as BSIM [68] and PTM [69] are based on physical model for rectangular devices. As the device is rectangular, the roughness along the gate width dimension is omitted, and the 3-dimensional device is abstracted to 2-dimensional with only X (along the gate length) and Z (vertical to the surface of gate). EGL, modeling one single parameter of above 2-dimensional compact models, is very difficult to fully catch effect of non-rectangular gate shape with compact device models for SPICE simulation.

In this chapter, I will propose a novel unified non-rectangular device compact model and circuit simulation methodology [71]. This is the first 3-dimensional post-lithographic device compact model. Different from previous works revaluing some specific parameter such as gate length, our work modifies the whole range of drain-source current model which is much more accurate than the two EGLs of ON and OFF states. Meanwhile, the impact of non-rectangular gate shape during parameter extraction of device model (BSIM or PTM) can be considered. Moreover, as an additional device modeling card, it is based on I-V properties of the devices, and thus it is compatible to current 2-D compact models and can be integrated with any existing device models

for post-lithography device/circuit simulation.

3.3 Preliminaries on Gate Slicing

In this section, the gate slicing method is presented. With the consideration of narrow width effect during gate slicing and current calculation, a new drain-source current can be calculated according to the current combination formula Eq. 3.1. Previous EGL methods are extracted from the new current under some specific working states as shown in last part of this section.

$$I_{ds}\left(SI, V_{ds}, V_{gs}\right) = \sum_{i=1}^{n} I_{ds,i}\left(L_i, \Delta W_i, V_{ds}, V_{gs}\right) \tag{3.1}$$

3.3.1 Slice the Gate

In the practical physics world, the manufactured gates are never rectangles with a uniform channel length and width, and with constantly scaling down, the non-rectangularity of gate shape is getting more and more impactive. As the channel width is much bigger than channel length in nanometer designs, the relative variations of channel width are much smaller than that of channel length. As [66], the equivalent channel width is calculated according to the same gate shape area at the width edge of the channel.

Previous LER papers [62, 63] have shown that slicing is a reliable method to simulate the channel length variations within a gate. And many post-OPC EGL papers [65–67] use this slicing method to study the impact of OPC.

After layout is designed and OPC is finished, the silicon image of each gate can be obtained as shown in Fig. 4.10(a). As the diffusion area is much bigger than the poly gate, the SI of the diffusion area is closer to rectangle after RET. Set W0 as the channel width. Then in the second step, the gate is sliced into small pieces vertical to the channel width direction (Fig. 4.10(b)). Each slice of the gate represents a single device with different channel lengths as shown in Fig. 4.10(c). In this way, one big non-rectangular gate is converted into a series of parallel small rectangular gates. There is some approximation during this process as the current flows at different gate width locations between source and drain are not paralleled to each other. With limited non-rectangularity of gate shape, the approximation is widely adopted [63, 65–67].

3.3.2 Narrow Width Effect

If the width of slice i in Fig. 4.10 is Δw_i, the length is L_i, the single device of slice i can be simulated by SPICE software. As the slicing process is just an artificial method, during the simulation of rectangular gate, L_i is a constant value of original gate length L_0 and the sum of Δw_i will the original width W_0. Slicing should have no impact on the device simulation results (Eq. 3.2).

$$I_{ds}\left(L_0, W_0, V_{ds}, V_{gs}\right) = \sum_{i=1}^{n} I_{ds,i}\left(L_i, w_i, V_{ds}, V_{gs}\right) \cdot \Delta w_i / W_0 \tag{3.2}$$

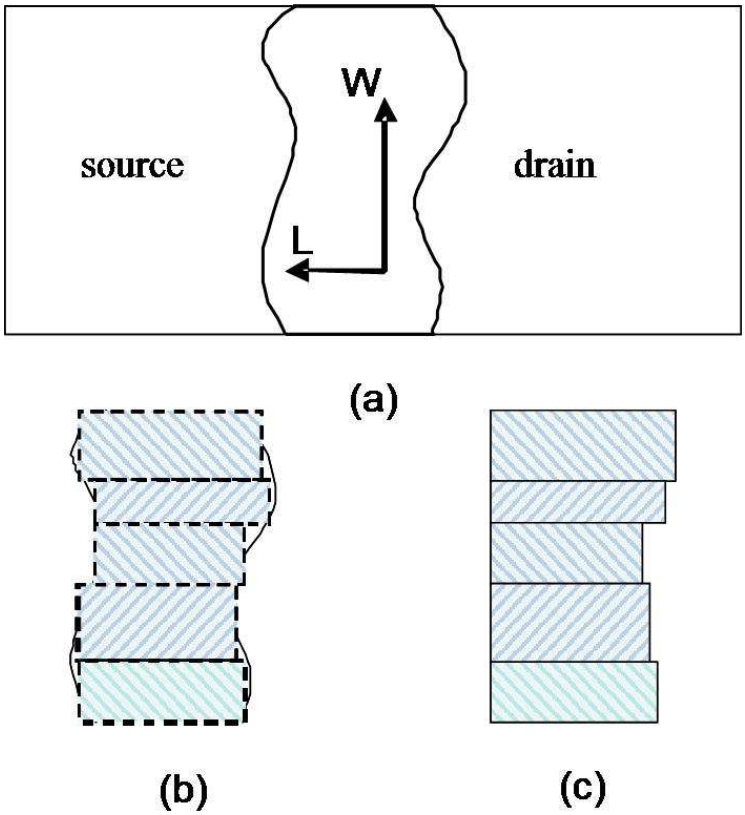


Figure 3.1: Slicing of a non-rectangle/non-uniform gate shape.
(a) The device with a non-rectangular gate; (b) Slice the non-rectangular into small pieces; (c) the equivalent gate is made up of rectangular pieces.

if,

$$\sum_{i=1}^{n} \Delta w_i = W_0$$
$$L_i \equiv L_0$$

.

Fig. 3.2 shows the IV curve of one NMOS with constant 65nm gate length and $2\mu m$ width based on device model of PTM [69]. The solid black

lines are direct SPICE simulation without the slicing, and the red dash lines are the sum of one hundred of parallel small devices with 20nm gate width after slicing. As the NMOS gate is rectangular, the curves of two simulations should be identical. However, with the impact of narrow width effect, the IV curves of the same device between with and without slicing are largely different. Without careful consideration, narrow width effect will accumulate over all device segments after slicing and cause serious error as shown in Fig. 3.2.

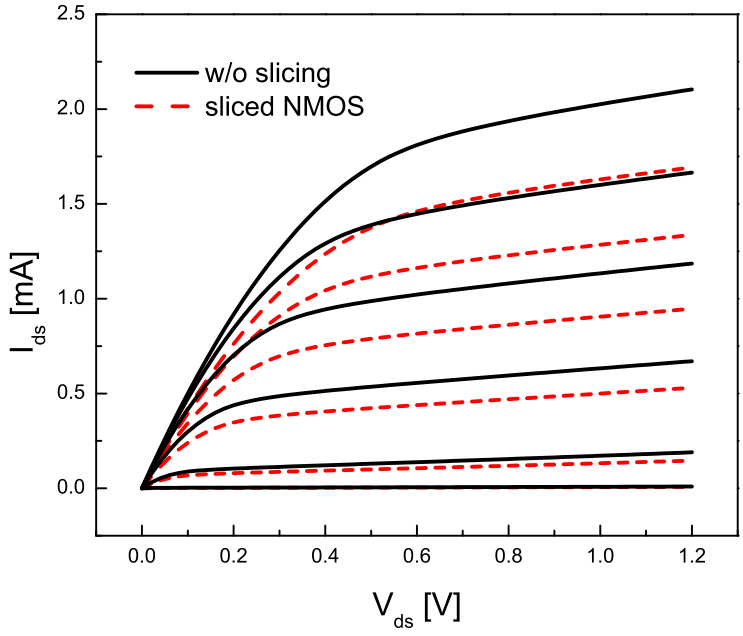


Figure 3.2: Impact of narrow width effect during slicing.

One way to eliminate the narrow width effect is to simulate two devices with width of $(W'+\Delta w_i)$ and W', and use the current difference for the current of the sliced device [66], where W' is a large device width. The disadvantage of this method is that when Δw_i is not small enough or constant W' is different

from gate width for simulation, the contribution of narrow width effect can not be as accurate as the original.

In this work, to accurately simulate the impact of narrow width effect, instead of widening the width of each slicing device to W', its original size W_0 is used. Then the drain-source current of each slice ($I_{ds,i}(L_i, \Delta w_i, V_{ds}, V_{gs})$) is calculated by $I_{ds,i}(L_i, W_0, V_{ds}, V_{gs}) \cdot \Delta w_i / W_0$. In this way, with slicing and enlarging of sliced gate width to the original size, the narrow width impact on current will be transformed back to the level before slicing.

$$\Delta I_{ds,i}(L_i, \Delta w_i, V_{gs}, V_{ds}) = \Delta I_{ds,i}(L_i, W_0, V_{gs}, V_{ds}) \cdot \Delta w_i / W_0 \tag{3.3}$$

3.3.3 Current Combination

After each slice is simulated according to its channel length L_i, width Δw_i, V_{gs}, V_{ds}, the drain-source current I_{ds} of the original gate with certain silicon image (SI) is calculated as Eq. 3.4

$$I_{ds}(SI, V_{ds}, V_{gs}) = \sum_{i=1}^{n} I_{ds,i}(L_i, W_0, V_{ds}, V_{gs}) \cdot \Delta W_i / W_0 \tag{3.4}$$

where,

$$\sum_{i=1}^{n} \Delta W_i / W_0 = 1$$

3.3.4 EGL and its Limitation

From Eq. 3.4, EGL (equivalent gate length) can be defined as the gate length function of V_{gs} and V_{ds} to keep the gate with the same drain-source

current under certain V_{gs} and V_{ds} as shown in below.

$$I_{ds}\left(L_{eq}, W_0, V_{ds}, V_{gs}\right) = I_{ds}\left(SI, V_{ds}, V_{gs}\right) \tag{3.5}$$

In most cases, EGL L_{eq} is not only dependent on the silicon image but also the device's working states (V_{gs} and V_{ds}). For a given SI, the EGL can be a function of V_{gs} and V_{ds} (Eq. 3.17).

$$L_{eq} = L_{eq}\left(V_{ds}, V_{gs}\right) \tag{3.6}$$

In [65–67], EGL for states ON and OFF are used to calculate the impact of non-rectangular gate with uniform gate length. An NMOS is ON when $V_{gs} > V_{th}$, and OFF when $V_{gs} < V_{th}$. And if the impact of V_{ds} on the drain-source current could be omitted, we can have the EGL of ON and OFF states separately.

$$\begin{array}{ll} I_{ds}\left(L_{eq,ON}, W_0\right) = I_{ds}\left(SI\right) & when\left(V_{gs} > V_{th}\right) \\ I_{ds}\left(L_{eq,OFF}, W_0\right) = I_{ds}\left(SI\right) & when\left(V_{gs} < V_{th}\right) \end{array}$$

From the above functions, we can see the limitation of previous EGL works: the impact of different values of V_{gs} and V_{ds} is simplified to only two states and so EGL method is only appropriate for some discrete cases from the continuous working states of a device. In the experiment section of this chapter, very inconsistent equivalent gate lengths will be presented to show the impact of different V_{gs} and V_{ds}. Moreover, in complicated circuit schematics, it is difficult to tell when to use ON or OFF EGL, especially in an automatic way.

3.4 A New Post-Litho Device Model

After reviewing previous EGL work for post lithography non-rectangular gates, a new post-litho device modeling card is developed directly from the drain-source current formula (Eq. 3.4) to the model of non-rectangular gates after lithographic process simulation. Different from previous works [65–67] which use EGLs to replace original values of uniform gate lengths in SPICE simulation, our post-litho device modeling card can more precisely simulate the impact of length variations within a single gate by continuously modifying the source/drain current. As it is extracted directly from the device I-V properties, it is independent on any device models and can be integrated with any existing device models.

3.4.1 Modeling the Difference between Non-Rectangular and Rectangular Devices

Most of existing compact model [68, 69, 72] are developed for gates of rectangular shapes. To make the new post-litho device model compatible to the existing compact modeling cards, the new non-rectangular model focuses on modeling the source/drain current difference between the non-rectangular and rectangular gates rather than the current of non-rectangular gates themselves.

If the current difference between rectangular segments with gate length L_i and L_0 is $\Delta I_{ds,i}\left(L_i, \Delta w_i, V_{gs}, V_{ds}\right)$, we can get the current difference between

non-rectangular and rectangular gates as Eq. 3.7 and Eq. 3.8.

$$\begin{aligned} \Delta I_{ds}\left(SI, V_{ds}, V_{gs}\right) &= I_{ds}\left(SI, V_{ds}, V_{gs}\right) - I_{ds}\left(L_0, W_0, V_{ds}, V_{gs}\right) \\ &= \sum_{i=1}^{n} \Delta I_{ds,i}\left(L_i, \Delta w_i, V_{gs}, V_{ds}\right) \cdot \Delta w_i \end{aligned} \tag{3.7}$$

As Eq. 3.3,

$$\begin{aligned} \Delta I_{ds}\left(SI, V_{ds}, V_{gs}\right) &= I_{ds}\left(SI, V_{ds}, V_{gs}\right) - I_{ds}\left(L_0, W_0, V_{ds}, V_{gs}\right) \\ &= \sum_{i=1}^{n} \Delta I_{ds,i}\left(L_i, W_0, V_{gs}, V_{ds}\right) \cdot \Delta w_i / W_0 \end{aligned} \tag{3.8}$$

Where, Δw_i is the width of the sliced segment, L_0 is the standard gate length of specific technology nodes (65nm, 45nm and etc.), and L_i is the gate length of the segment with is close but different from L_0.

Now, if known the current difference of sliced segments between different gate lengths, the total current difference between rectangular gates and non-rectangular gates would be available. In the next part, different ways to model and calculate the segment current difference would be presented.

3.4.2 Current Model through Sliced Rectangular Segment

In Eq. 3.8, to calculate the I-V curves $I_{ds,i}(SI, V_{ds}, V_{gs})$ of any silicon image under certain gate-source and drain-source voltage, the current $I_{ds,i}(L_i, W_0, V_{ds}, V_{gs})$ of certain channel length L_i and width W_0 should be known. Thus, current $I_{ds,i}(L_i, W_0, V_{ds}, V_{gs})$ are precalculated and collected in our model. There are mainly three ways to extract the current modification function [73]: table look-up, analytical model, empirical model.

3.4.2.1 Table Look-up

One of the most straightforward way to get values of $I_{ds,i}(L_i, W_0, V_{ds}, V_{gs})$ is to do the simulation or test and record the results in a table. As the current is a function of many parameters including gate length, width, drain-source voltage, and gate-source voltage, the whole table will be very big and multi-dimensional SPLINE approximation is necessary to keep the function curves smooth.

3.4.2.2 Continuous I_d Model for Short Channel MOSFET

Another option is to use analytical model. The continuous drain current Eq. 3.9 [73] is used to model the relationship of I_{ds} and gate length, gate width, V_{gs} and V_{ds}. The advantage of this short channel effect model is that the current and its first derivatives are continuous at the segmentation points of general segmented functions, while the MOSFET models from level 1 to level 4 in SPICE are divided into 3 different regions (sub-threshold region, linear region and saturation region), where the first derivatives of the current are not continuous. If the parameters in the function are available or could be easily extracted, $I_{ds,i}(L_i, W_0, V_{ds}, V_{gs})$ can be calculated by Eq. 3.9 directly.

$$I_{ds}(L) = \frac{W\mu_s C_{ox}\left(V_{gsx} - V_{th} - 0.5\alpha V_{dsx}\right)V_{dsx}}{\left(L - l_d\right)\left(1 + V_{dsx}/\left(L - l_d\right)\mathrm{E}_c\right)} \tag{3.9}$$

where,

$$V_{gsx} = \eta V_t \ln\left\{1 + \exp\left[\left(V_{gs} - V_{th}\right)/\eta V_t\right]\right\} + V_{th}$$

$$V_{dsx} = V_{dsat}\left\{1 - \frac{1}{B}\ln\left[1 + \exp\left(A\left(1 - V_{ds}/V_{dst}\right)\right)\right]\right\}$$

$$V_{dsat} = \frac{(1-\delta_0)(V_{gs}-V_{th}) + \alpha L E_c}{\alpha(1-2\delta_0)} \cdot \left[\sqrt{1 + \frac{2\alpha(V_{gs}-V_{th})LE_c(2\delta_0-1)}{((1-\delta_0)(V_{gs}-V_{th})+\alpha L E_c)^2}} - 1\right]$$

Therefor, the current difference of two rectangular gate segments with length L_i and L_0 would be:

$$\begin{aligned}
&\Delta I_{ds,i}(L_i, W_0, V_{gs}, V_{ds}) \\
=&I_{ds}(L_i, W_0, V_{gs}, V_{ds}) - I_{ds}(L_0, W_0, V_{gs}, V_{ds}) \\
=&W_0\mu_s C_{ox} \cdot \frac{(V_{gsx} - V_{th} - 0.5\alpha V_{dsx}(L_i)) \cdot V_{dsx}(L_i)}{(L_i - l_d) + V_{dsx}(L_i)/\mathrm{E}_c} \\
-&W_0\mu_s C_{ox} \cdot \frac{(V_{gsx} - V_{th} - 0.5\alpha V_{dsx}(L_0)) \cdot V_{dsx}(L_0)}{(L_0 - l_d) + V_{dsx}(L_0)/\mathrm{E}_c}
\end{aligned} \tag{3.10}$$

The notations of variables in Eq. 3.9 are shown in Table 3.1.

However, Eq. 3.10 is very complicated and it is lack of the parameter extractor to convert the BSIM or PTM model parameters into parameters in Table 3.1. As a result in the rest of this work, an empirical drain current model is adopted.

3.4.2.3 Empirical Drain Current Fitting Model

To reduce the data space storing look-up table, curve fitting with empirical function is applied. According to our I-L curve fitting experience, Eq. 3.11 well fits the drain/source current-length curves at different V_{gs} and V_{ds} with different values of three fitting parameters a, b, and c.

$$I_{ds} = \exp\left(a + b \cdot L + c \cdot L^2\right) \tag{3.11}$$

Table 3.1: Variable notations of continuous I_d.

Symbol	Notation
I_{ds}	Drain Current
W	Gate width
L	Gate length
l_d	Length near the drain end due to channel length modulation
V_{ds}	Drain-source voltage
V_{dsx}	Effective drain-source voltage
V_{dsat}	Drain saturation voltage
A	Parameter. Large value of A yields steep transitions between the linear and saturation regions while small values result in smooth transitions.
V_{gs}	Gate voltage
V_{gsx}	Effective gate voltage
V_{th}	Threshold voltage
η	Parameter. Physically, it signifies the capacitive coupling between the gate and silicon surface.
μ_s	MOSFET surface mobility
C_{ox}	Gate oxide capacitance per unit area
α	Body factor term
E_c	Critical field for the carrier velocity saturation [V/cm]
δ_0	A function of the electrical field

Though the database of table look-up can be greatly reduced, the three parameters are not consistent when V_{gs} and V_{ds} change. So ΔI_{ds} is calculated and stored for the consistency purpose during interpolation.

$$\Delta I_{ds,i}\left(L_i, W_0, V_{gs}, V_{ds}\right) = \exp\left(a + bL_i + c{L_i}^2\right) - \exp\left(a + bL_0 + cL_0^2\right) \quad (3.12)$$

3.4.3 Impact of Parameter Extraction

The parameters of device models such as BSIM3 [68], BSIM4 [68], and PSP [72] are extracted from the experimental results of real manufactured devices. In addition to the models, the parameter extraction process is also provided together [68]. As we discussed in the introduction section, in the practical physical world as the printability of lithographic patterns is greatly challenged, the manufactured devices could not be perfectly rectangular any more. Therefore the impact of some specific non-rectangular gate shape on the device electrical performance has already been involved in a huge number of device models parameters. Only having Eq. 3.8 is not enough to fully model the non-rectangular devices and we also need parameter extraction of the manufactured non-rectangular gates.

Previous EGL methods [65–67] always assume that those device models (BSIM or PTM) are extracted from perfect rectangular gate with constant channel lengths. This assumption is not reliable and will overestimate the impact of channel length variations within a gate. Our post-litho modeling card presented in this section would show how to extract the rectangular device model from measurement data of non-rectangular gates and model the precise

performance difference between devices of two non-rectangular SI. Thus non-rectangular SI information of the device models can be well considered in this modeling card.

In this part, a parameter extraction solution for non-rectangular gates will be presented. As the flow and tools of parameter extraction for existing rectangular models (e.g. BSIM [68]) are mature and widely used, the parameter extraction of non-rectangular gates should be compatible to existing parameter extraction solution and reuse it for non-rectangular gates.

3.4.3.1 An Example of Non-rectangular Device Modeling

As mentioned before, the parameter extraction of BSIM or other device modeling has already considered the impact of certain non-rectangle gate on device performance indirectly, as those parameters are extracted from the real manufactured devices. A big number of parameters in those device models are used to fit the difference between experimental data and the theoretical models. And finally, these device models can fit the corresponding practical non-rectangular devices very well through parameter extraction.

In Fig. 3.3, the silicon images of two manufactured devices are shown. Obviously, they are non-rectangular, then device models and their parameters are extracted from the devices. If the devices in the circuit have absolutely the same silicon image, then existing device models would give the right performance prediction. Yet if assuming perfect rectangular gate of device model parameter extraction, after EGL [65–67], a new gate length will be calculated

which will induce errors in device performance estimation of non-rectangular gates.

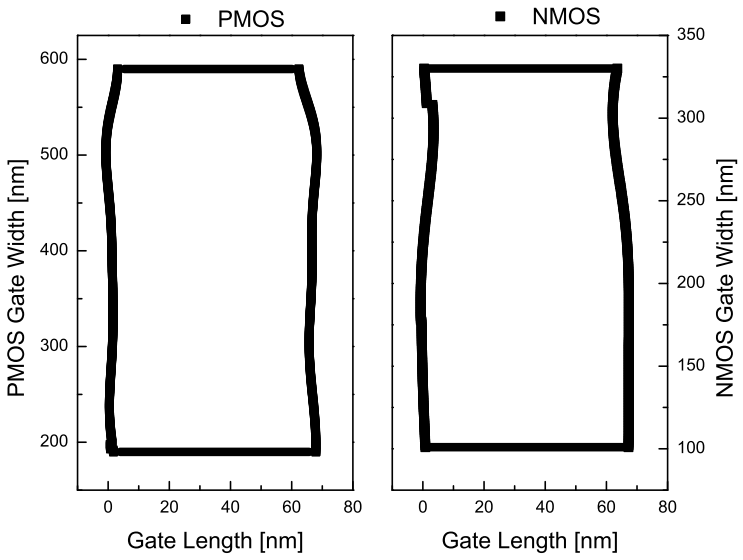


Figure 3.3: Silicon Image of the device for model and parameter extraction.

After getting device models from certain non-rectangular gates (such as device in Fig. 3.3), during the post-lithographic circuit simulation, a series of different non-rectangular gate modeling cards in the circuit are necessary. Thus the whole process is to do non-rectangular device simulations based on device models of non-rectangular device parameter extraction from another non-rectangular shape, and the right post-OPC device characterization should transform device model of specific non-rectangular gates (such as BSIM with parameters supplied by fabs) into the device model of a different general non-rectangle gate, which varies because of different RET/OPC strategies, different layout environments of circuits and also different manufacturing

process variations.

3.4.3.2 Gate Slicing Reordering & Parameter Extraction

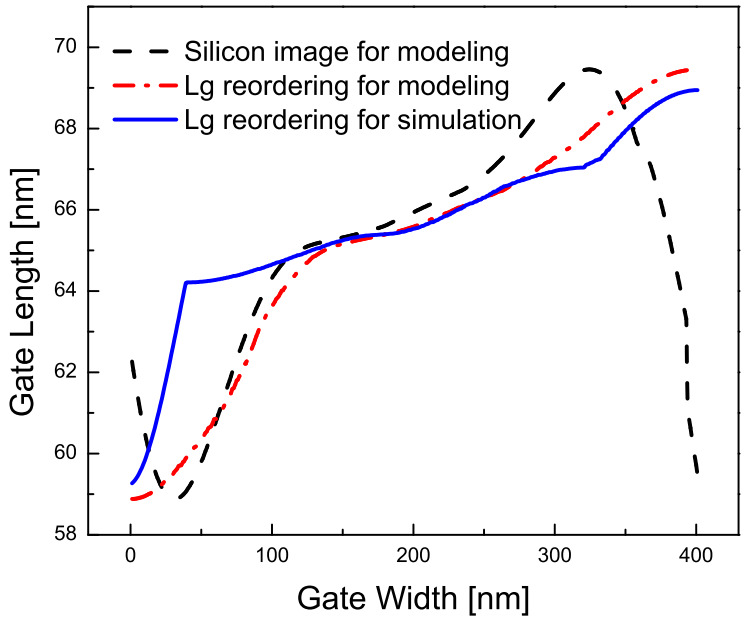


Figure 3.4: Gate slicing reordering.
The black line is the SI for modeling and parameter extraction; the red line is the reordered black line. The blue line is the gate length reorder of gate SI for circuit simulation.

The segments of gate sizing in Fig. 4.10 can be reordered according to the gate length in each segment if the current in each segment is paralleled to each other. The reordering of gate length of PMOS silicon image is shown in Fig. 3.4. The black dash line is the gate length without reordering (relative to (c) in Fig. 4.10) for the manufactured device in order to extract model parameter. The red dash dot line is reordered black line. If the blue solid line is the reordered gate length for circuit simulation with a different silicon image,

the impact on device performance should come from the gate length difference between the blue solid line and the red dash line. The previous works of EGL [65–67] do not consider the non-rectangular gate shape during device model parameter extraction. Those works just assume that the devices for parameter extraction are rectangular and set red dash line to be a constant gate length (such as $65nm$). Such difference between blue solid line and red dash dot line is magnified and the impact of non-rectangular gate shape is overestimated.

3.4.3.3 Device Model for Rectangular Gate with Extraction Involved

After pointing out the shortcoming of existing EGL works, a device model for rectangular gates is necessary to apply Eq. 3.8. After applying slicing method on the non-rectangular gate for device model and its parameter extraction again, the device model for rectangular gate can be deduced. If SI_{EX} is the silicon image of the gate for device model parameter extraction,

$$\begin{aligned} &\Delta I_{ds,EX}\left(SI_{EX}, V_{ds}, V_{gs}\right) \\ =&I_{ds}\left(SI_{EX}, V_{ds}, V_{gs}\right) - I_{ds}\left(L_0, W_0, V_{ds}, V_{gs}\right) \\ =&I_{ds,\,\mathrm{mod}\,el}\left(L_o, W_0, V_{ds}, V_{gs}\right) - I_{ds}\left(L_0, W_0, V_{ds}, V_{gs}\right) \end{aligned} \tag{3.13}$$

where $\Delta I_{ds,EX}$ is the current difference between non-rectangular gate for parameter extraction and $\Delta I_{ds,model}$ is the current directly calculated from device models and their parameters.

Eq. 3.13 gives the current difference between gate with SI_{EX} and a corresponding rectangle device. So the current of ideal rectangle device can

be calculated by subtracting $I_{ds,EX}$ from current simulation results of SPICE based on BSIM or PTM model and Eq. 3.15 would provide the current of perfect rectangular gates.

$$\begin{aligned} & I_{ds}\left(L_0, W_0, V_{ds}, V_{gs}\right) \\ =& I_{ds,\,\mathrm{mod}\,el}\left(L_o, W_0, V_{ds}, V_{gs}\right) - \Delta I_{ds,EX}\left(SI_{EX}, V_{ds}, V_{gs}\right) \end{aligned} \tag{3.14}$$

3.4.4 Post-Litho Device Modeling Card

From Eq. 3.8 and Eq. 3.15, the drain-source current difference of post-litho device can be calculated as the post-litho modeling card as shown in

$$\begin{aligned} & \Delta I_{ds,litho}\left(SI, V_{ds}, V_{gs}\right) \\ =& I_{ds}\left(SI, V_{ds}, V_{gs}\right) - I_{ds,\,\mathrm{mod}\,el}\left(L_o, W_0, V_{ds}, V_{gs}\right) \\ =& \Delta I_{ds}\left(SI, V_{ds}, V_{gs}\right) - \Delta I_{ds,EX}\left(SI_{EX}, V_{ds}, V_{gs}\right) \end{aligned} \tag{3.15}$$

And it will be added to each MOSFET with any existing device model to modify the drain/source current by a value of $\Delta I_{ds,litho}\left(SI, V_{ds}, V_{gs}\right)$ with the SI information as the input arguments. Each MOSFET will be replaced by a post-litho module made up of their original MOSFET with uniform gate length and an additional post-litho modeling card as shown in Fig. 3.5.

3.4.5 Post-Litho Circuit Simulation Flow

The circuit/cell level simulation flow is proposed based on the above post-litho device modeling card. Instead of two values of equivalent gate lengths, the post-litho modeling card modified the I-V curve of a single device under certain silicon image with continuous working states (under different V_{gs}

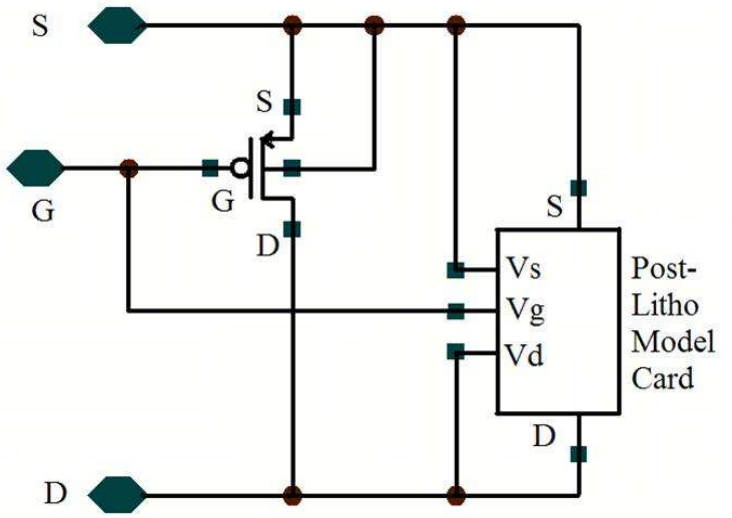


Figure 3.5: Module Schematic of Post-litho Device Modeling Card.

and V_{ds}), and thus much more cell level simulation can be performed based on the modeling card.

As the process variations will change the silicon image of devices, the impact of lithographic relative process variations (such as defocus and variations of dose) on the circuit electrical performance can also be precisely simulated based on the post-litho device modeling card.

3.4.5.1 Simulation Flow on Circuit Level

The design flow of Integrated Circuit, especially digital IC is able to be divided into different levels, and at each design level, the internal details can be abstracted away with the replacement of model which could be regarded as block box to the designers or design tools from other levels. This hierarchical property is believed to be the key ingredient for the success of digital circuit

design [74]. Unfortunately, with the continuously scaling down, more and more detailed information has to be fed back from the bottom level of design flow to earlier levels. Such across-design-level approach includes research on Design for Manufacturability, wire length prediction logic synthesis and so on. Over the whole design and manufacture flow, the post layout simulation is extremely important as it is the last step before both taping out and the great economic gap between manufacturing and simulation cost.

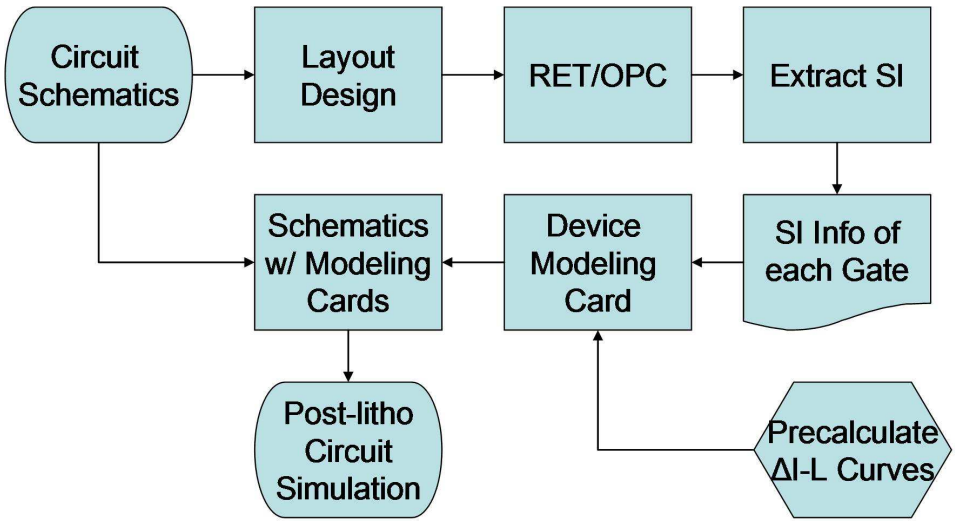


Figure 3.6: Post-litho circuit simulation flow.

Fig. 3.6 shows the post-litho circuit simulation flow of this work. The ΔI-L of devices from Eq. 3.12 and Eq. 3.15 is precalculated and stored in the post-litho device modeling card. The SI information of each gate is input as the arguments to the modeling card, and each MOSFET in original circuit schematics is modified by the corresponding modeling card (as shown in Fig. 3.6). Post-litho circuit simulation is performed based on the new schemat-

ics.

3.4.5.2 Lithographic Process Variations

The lithographic process variations will change the silicon image, which could be simulated by lithographic software. And the impact of lithographic variation on the whole circuits can be simulated in our post-litho circuit simulation flow when SI with lithographic variations is input into our post-litho device modeling card. According to [75], the relationship between edge placement error (EPE) of the gate and lithographic process variations (include dose I_{th} variation and defocus z) can be expressed approximately as Eq. 3.16.

$$E\left(I_{th}, z\right) = E_{iso} + a_0\left(1 + a_1 z^2\right)\left(I_{th} - I_{th,iso}\right) \tag{3.16}$$

where E_{iso}, a_0, a_1 and $I_{th,iso}$ can be regarded as constant during the process variation analysis [75].

The gate length variation within gate can be calculated based on EPE and misalignment [67]. As only the second order of defocus z appears in Eq. 3.16, no matter positive or negative of defocus will have the consistent impact (increase or decrease) on EPE and gate length. For the quadratic relationship between EPE/gate length and depth of focus, and random symmetric variation will cause systematic asymmetric variations of gate length. In our simulation cases of the following results section, any defocus will decrease the gate length, and in consequence controllable level of defocus will be helpful to reduce gate length, push the technology node forward and increase device

timing performance. As shown in Eq. 3.16 and [75], the dose variation has linear relation with EPE and gate length, this process variation has almost no contribution to the expectation of the gate length.

3.5 Experimental Results of Post-Litho Simulation

The simulation results are shown in this section. The simulation environment is explained first. Then the equivalent gate length of different device working states is studied. After the post-litho device modeling card is validated under special working states, timing (delay and slew) and power (dynamic and static) are discussed. Finally, inverter chain is simulated to compare EGL and the proposed post-litho modeling card.

3.5.1 Introduction of Testing Cases

According to the design rules of 65nm, the inverter layout is drawn and OPC as well as lithographic simulations are finished. During the simulation, the process variations are considered and the statistical expectation is used to investigate the impact of process variation on the mean value of gate lengths. In the following experiments, we will show that effect of process variations will lead to big difference of the performance expectations of the devices and circuit cells.

The gate length is 65nm, and gate width of PMOS is 400nm while gate width of NMOS is 200nm. The 65nm gate length on the drawn layout does not mean the silicon image will be rectangular. If we set gate length

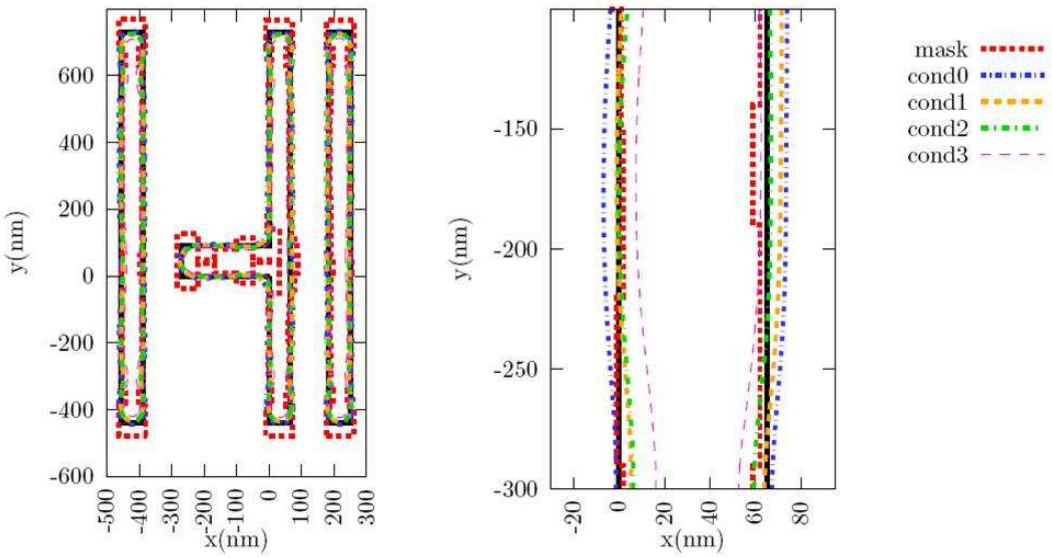


Figure 3.7: Layout and post-OPC simulation of 65nm inverter.
The left one is the whole view of the pattern, while the right one is NMOS region of the inverter.
cond0: $I_{th} = 0.143$, $z = 0$; cond1: $I_{th} = 0.143$, $z = 80nm$; cond2: $I_{th} = 0.157$, $z = 0$; cond3: $I_{th} = 0.157$, $z = 80nm$;

to be 65nm in device model, the properties of the device should be correctly simulated. Therefore we set the corresponding silicon image of the gate for the device model to some non-rectangular shape as shown in Fig. 3.3. As mentioned previously, the non-rectangular device model should characterize the difference between device in circuit cell with non-rectangular silicon image (as shown in Fig. 3.7) and the non-rectangular device (in Fig. 3.3) during parameter extraction.

3.5.2 Uncertainty of Equivalent Gate Length

As mentioned above, at different V_{gs} and V_{ds}, the modification of drain-source current in Eq. 3.8 is different. As the equivalent gate length is calculated from the current, EGL of those continuously changing V_{gs} and V_{ds} should be various. As a function of V_{gs} and V_{ds}, more accurate equivalent gate lengths are calculated according to Eq. 3.17.

$$I_{ds}\left(L_{eq}\left(V_{gs},V_{ds}\right),W_0,V_{ds},V_{gs}\right)=I_{ds}\left(SI,V_{ds},V_{gs}\right) \tag{3.17}$$

where L_{eq} is the equivalent gate length, and it depends on the working state of V_{gs} and V_{ds}.

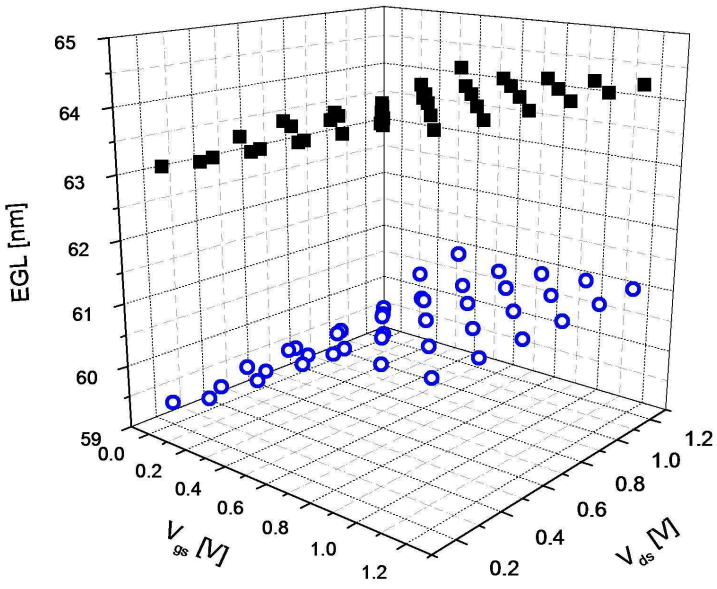


Figure 3.8: Different equivalent gate lengths of NMOS.
Black square does not consider process variations while blue circles are with process variations.

Fig. 3.8 shows that the ranges of EGLs of NMOS are various at different

V_{gs}, and even for the same V_{gs}, different values of V_{ds} also induce different EGLs. EGL of PMOS is shown in Fig. 3.9.

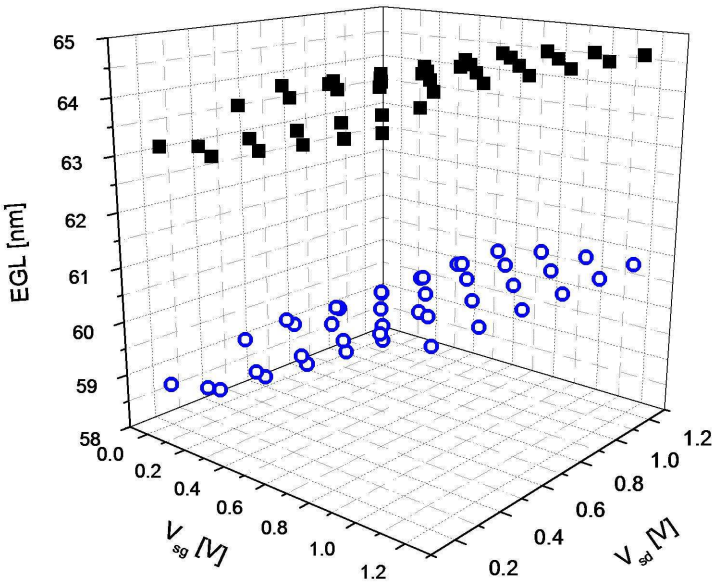


Figure 3.9: Different equivalent gate lengths of PMOS.
Black square does not consider process variations while blue circles are with process variations.

After considering the lithographic variations (defocus and dose variations), the expectation of the gate shape will be different and the calculated EGL will be changed obviously. According to our simulation cases, the lithographic variations of defocus will induce the decrease of EGL.

If the points in Fig. 3.8 are projected to "EGL" and V_{gs} plane, candle stick pattern in Fig. 3.10 can be generated. We can see EGL changes with gate-source voltage, and even for the same V_gs, at different V_ds, EGL will differ from each other. When the transistor is ON, EGL is slightly dependent on V_{gs}

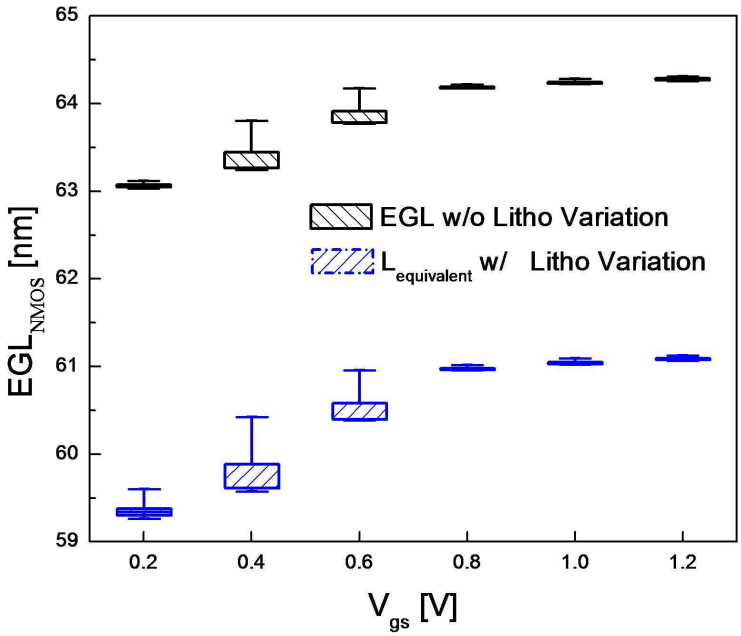


Figure 3.10: Candle stick pattern of different NMOS EGL.
In each pattern, the top line is the maximum value, the bottom line is the minimum value, and the box in the middle is the range of standard error with coefficient 1.

and almost constant when V_{ds} changes. Even though the dependence on V_{ds} is not so strong as that on V_{gs}, the dependence on V_{ds} is obvious when V_{gs} is in the middle range and the gate is partially on, which is omitted in [76].

Previous works of equivalent gate length for each device only consider two values: $L_{eq,on}$ for ON state and and $L_{eq,off}$ for OFF state. As the state ON and OFF of a device are decided by whether V_{gs} is bigger than V_{th}, from the above results we can see that $L_{eq,on}$ and $L_{eq,off}$ are rough estimation for NMOS and PMOS as there is such a big range of equivalent gate length variations under various V_{gs} and V_{ds}. 2 ON/OFF values of EGL are not enough for

accurate simulation.

3.5.3 Validation of Post-Litho Device Model

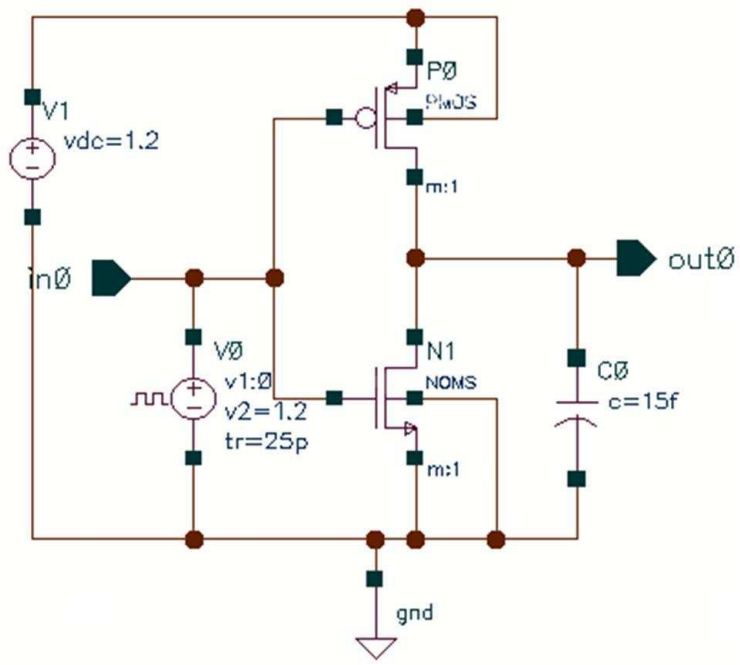


Figure 3.11: Schematic of inverter cell.

The post-litho device module (Fig. 3.5) is validated in the circuit level simulation. Fig. 3.11 is the schematic of an inverter cell circuit. In Fig. 3.12, the inverter with post-litho non-rectangular gates is simulated to compare voltage, current, timing and power issues of post-litho modeling card with EGL methods. The input is a pulse voltage between 0V and 1.2V with 2ns period and 25ps rising and falling time. The load capacitance is 15fF. Note that to verify the module of modeling card, the modeling card in this section does not involve the impact of the non-rectangle gate shape/SI of device models

which is not considered by EGL method.

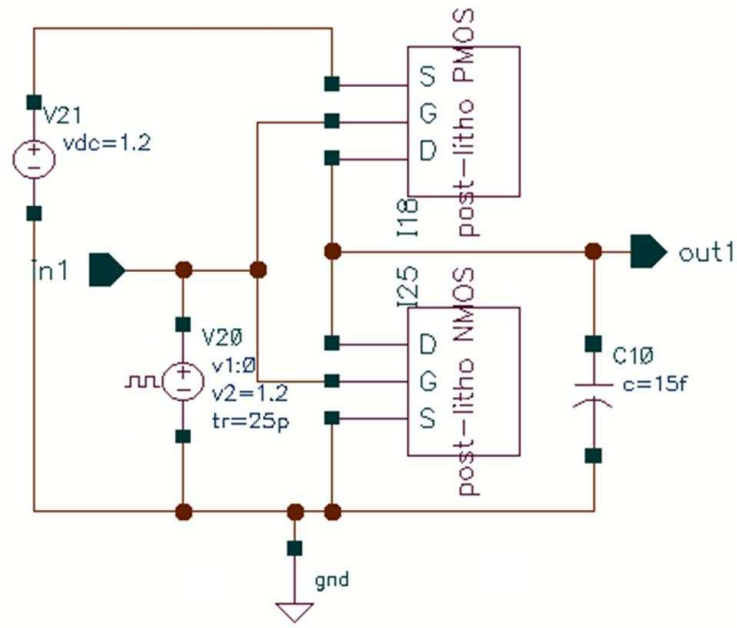


Figure 3.12: Schematic of inverter cell for Post-litho simulation.

As mentioned before, the EGL model is valid when the device is completely on or off. Therefore, timing and leakage results of the post-litho modeling card for non-rectangular gates is compared with EGL model. As there is no parameter extraction consideration in EGL model, the parameter extraction from non-rectangular gates is not involved in the post-litho non-rectangular gates.

From Fig. 3.13, the device modeling card module (red solid lines) is close to the V_{out} points simulated by ON equivalent gate length, especially at middle voltage value of 0.6V which is used to define the circuit delay. The delay data of different methods can be found in Table 3.2 and slew results in

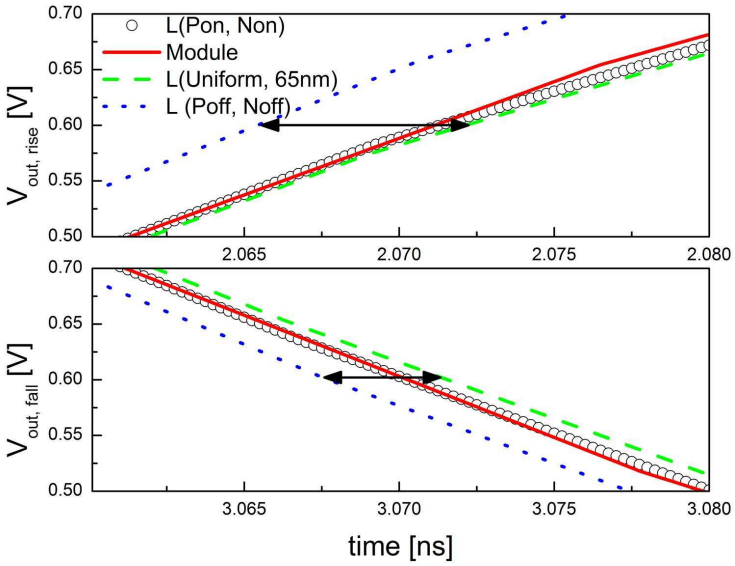


Figure 3.13: Comparison of rising and falling timing for validation.
The green dash lines are simulation results from rectangular gates with uniform $65\mu m$ gate length. The red solid lines are from post-litho non-rectangular modeling card. The black circles are from simulation by setting both NMOS and PMOS to $L_{eq,ON}$. The blue dots are from simulation of $L_{eq,OFF}$.

Table 3.3.

Table 3.2 gives output delay comparison between different methods. "ON EGL" means that the gate lengths of devices in the cell schematics are set to be their ON equivalent gate length, while "OFF EGL" means all devices use their OFF EGL. "L=65nm" is that the gate lengths of devices 65nm which is the uniform gate length on the layout. "Post-litho Model" means the results based on our proposed post-litho unified non-rectangular device and circuit simulation method. However, to validate the post-litho model, here the parameter extraction and SI_{EX} are not considered, as EGL methods do

Table 3.2: Post-litho model validation by inverter output delay.

	Rising Delay		Falling Delay	
	(ps)	diff.	(ps)	diff.
ON EGL	58.8	-	57.8	-
L=65nm	59.6	1.42 %	59.1	2.21 %
OFF EGL	52.6	-10.5 %	55.2	-4.43 %
Post-litho Model	58.6	-0.28 %	57.8	0.01 %

not use it. According to Table 3.2, the difference between "ON EGL" and "Post-litho model" is $5-50X$ smaller than the difference between "ON EGL" and other methods ("L=65nm" or "OFF EGL").

Table 3.3: Post-litho model validation by inverter output slew.

	Rising Slew		Falling Slew	
	(ps)	diff.	(ps)	diff.
ON EGL	125	-	108	-
L=65nm	126	0.94 %	110	2.09 %
OFF EGL	113	-9.53 %	103	-4.27 %
Post-litho Model	126	0.37 %	108	0.34 %

Table 3.3 gives output delay comparison between different methods. The difference between "ON EGL" and "Post-litho model" is $2.5-25X$ smaller. However, the above simulation results do not conclude that "ON EGL" is more accurate than "post-litho model". In the timing simulation, both NMOS and PMOS are not always ON. "Post-litho model" should have more accuracy than "ON EGL" method. Anyway, as "ON EGL" is much more accurate/reasonable than "L=65nm" and "OFF EGL", the smaller difference

between "Post-litho model" and "ON EGL" in delay and slew validate the proposed non-rectangular post-litho device modeling card (Fig. 3.5) and circuit simulation method (Fig. 3.6).

Leakage simulation is also used in the validation. As the simulation of leakage is static and the voltages on PMOS and NMOS would not be changed, it is possible to figure out what the right work state is and what exact EGL should be used for each transistor in the inverter circuit. Fig. 3.14 is the leakage current through PMOS drain when PMOS is completely OFF and NMOS is completely ON.

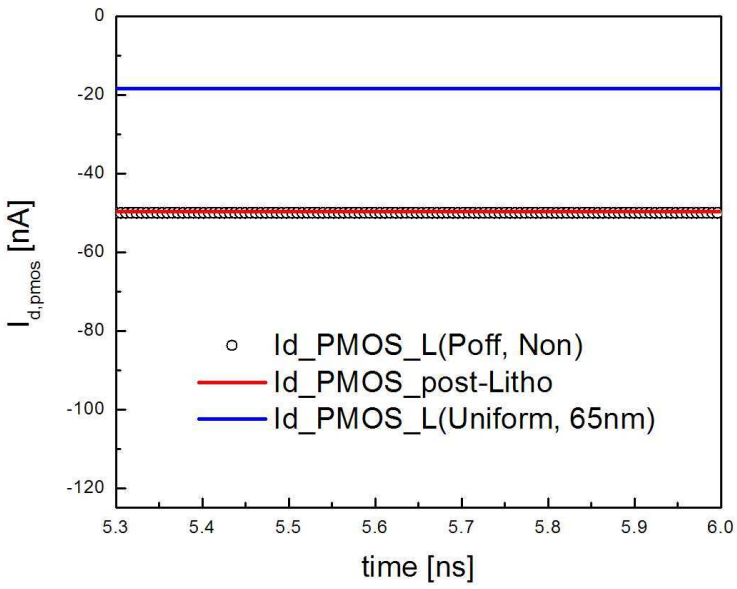


Figure 3.14: Constant leakage current through drain of PMOS.

As the state of each device in the cell is known, the right simulation (circles in the figure) with EGL method should set PMOS gate length to be

"OFF EGL" and NMOS gate length to be "ON EGL". The simulation of uniform 65nm gate length is the blue line in Fig. 3.14 and red line is from "post-litho model". The drain current of post-litho model can well overlap the points of PMOS drain current of equivalent gate length in OFF state.

According to above simulation, our post-litho model is validated and will be regard as "golden module" for future simulation. In the rest parts of this section, the non-rectangular gate shapes SI_{EX} for model parameter extraction will be involved into the post-litho non-rectangular device model.

3.5.4 Timing Results of Post-Litho Circuit Simulations

After validating the unified post-litho non-rectangular device modeling card, complete modeling cards (with the consideration of non-rectangular SI_{EX} for model parameter extraction) are used for timing analysis of post-litho circuit simulation.

Fig. 3.15 shows the rising/falling delay of ON EGL method, our unified model without and with variations. The light arrows show the delay from V_{in} to V_{out}, and the black arrows are for the impact of lithographic process variations. Here, "w/ Variations" means the statistical expectation based on post-litho non-rectangular model with the consideration of normal distribution of dose and defocus variations [42, 75],.

Delay (Table 3.4) and slew (Table 3.5) of different simulation methods and conditions are compared.

According to Table 3.4 and 3.5 the equivalent gate length method has

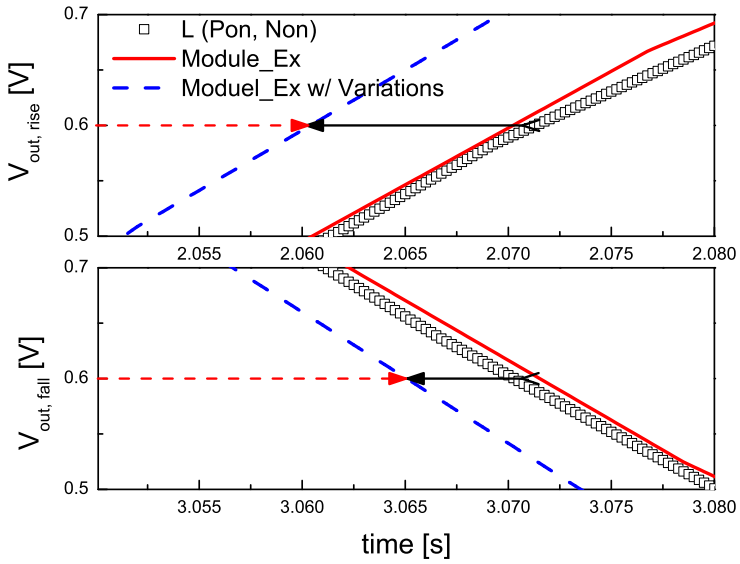


Figure 3.15: Rising and falling timing of V_{out}.

Table 3.4: Output delay comparison.

	Rising Delay		Falling Delay	
	(ps)	diff.	(ps)	diff.
Post-litho Model w/o Variations	57.7	-	59	-
ON EGL w/o Variations	58.8	1.87 %	57.8	-2.09 %
Post-litho Model w/ Variations	47.9	-17.0 %	52.6	-11.0 %

about 2% underestimation for rising delay and 2% overestimation for falling delay. Both rising and falling slews are overestimated by EGL. After considering SI_{EX} during device modeling parameter extraction, ON EGL methods for delay could be very inaccurate and may mislead the design optimization.

As the same litho variation has different impact on PMOS and NMOS, its impact on timing issues of rising and falling are different, which is also

Table 3.5: Output slew comparison.

	Rising Slew		Falling Slew	
	(ps)	diff.	(ps)	diff.
Post-litho Model w/o Variations	124	-	110	-
ON EGL w/o Variations	125	1.18 %	108	2.07 %
Post-litho Model w/ Variations	104	-16.2 %	99.7	-9.47 %

shown in Table 3.4 and 3.5 with the title of "Post-litho Model w/ Variations" meaning that the proposed post-litho model is used to simulate the impact of lithographic variation.

3.5.5 Power Dissipated on the Post-Litho Cell

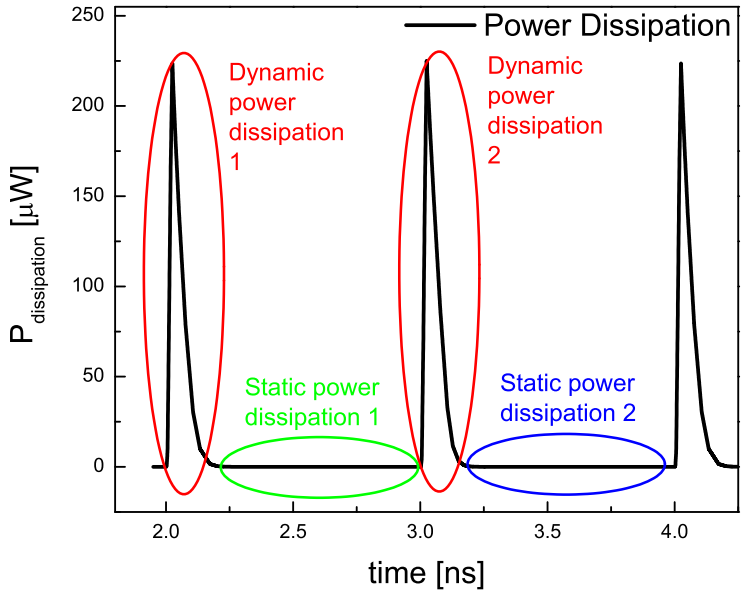


Figure 3.16: Power dissipation on the inverter.

Since dissipated power of the whole circuit not only directly affect power

consuming, but also leads to thermal issues and elevates the temperature of the whole chip, power dissipation of the whole cell rather than the single device is studied in this part. As the PMOS and NMOS alternatively play the main character role in the inverter cell, there are 2 dynamic zones and 2 static zones for power consuming analysis as shown in Fig. 3.16.

The comparison results are shown in Table 3.6. It compares the mean as well as peak dynamic power dissipation of two zones on the cell with different simulation methods and conditions.

Table 3.6: Dynamic power dissipation comparison.

Dynamic - 1	Mean		Max	
	(μW)	diff.	(μW)	diff.
Post-litho Model	41.13	-	223.7	-
EGL ON	42.68	3.79 %	227.6	1.78 %
EGL OFF	42.6	3.59 %	247.6	10.70 %
Post-litho Model w/ Variations	33.56	-18.40 %	211.9	-5.30 %
Dynamic - 2	Mean		Max	
	(μW)	diff.	(μW)	diff.
Post-litho Model	42.73	-	225	-
EGL ON	42.73	0.01 %	229	1.79 %
EGL OFF	42.74	0.02 %	236.9	5.29 %
Post-litho Model w/ Variations	37.73	-11.70 %	219.3	-2.60 %

The static power dissipation is shown in Table 3.7. It is impressive that the static power/leakage simulation of "EGL OFF" may be seriously wrong (up to $1.5X$ overestimation). This is because the EGL for OFF state does not consider the fact that the devices for parameter extraction are also

non-rectangular gates. Without taking into account the non-rectangular gate shapes for model parameter extraction, EGL methods could seriously overestimate leakage even though OFF EGL is used for leakage simulation. There are also obvious errors in other issues by any EGLs such as dynamic and static power consuming (Table 3.6 and 3.7).

Table 3.7: Static power dissipation comparison.

	Static 1		Static 2	
	(nW)	diff.	(nW)	diff.
Post-litho Model	24.3	-	29.2	-
EGL ON	26.5	9.05 %	31.1	6.67 %
EGL OFF	62.2	156 %	36.6	25.30 %
Post-litho Model w/ Variations	24.8	2.11 %	29.2	-0.12 %

Note that in Table 3.6 and 3.7, "Post-litho Model w/ Variations" means the statistical expectation based on normal distributed defocus and dose variations. As dose variation has linear impact on gate length [75], the shift of statistical expectation is caused by defocus variation.

The cell level simulation in Table 3.6 and 3.7 shows that certain litho variation (defocus) is helpful to reduce total power consuming as average dynamic power can decrease by 10% to 20%, while static power only increases by just 0-2%. This is coincident to the common sense of gate length shrinkage. According to (11), depth-of-focus variations (no matter z is positive or negative) will lead to the decrease of gate length in this simulation case.

3.5.6 Power Supply Current Simulation

The requirement of current supply from voltage source V_{dd} will directly affect the supply voltage level and power distribution of the whole chip, and may challenge the integrity of the power distribution network [77], so the current/power supplied by voltage source is studied separately in this part.

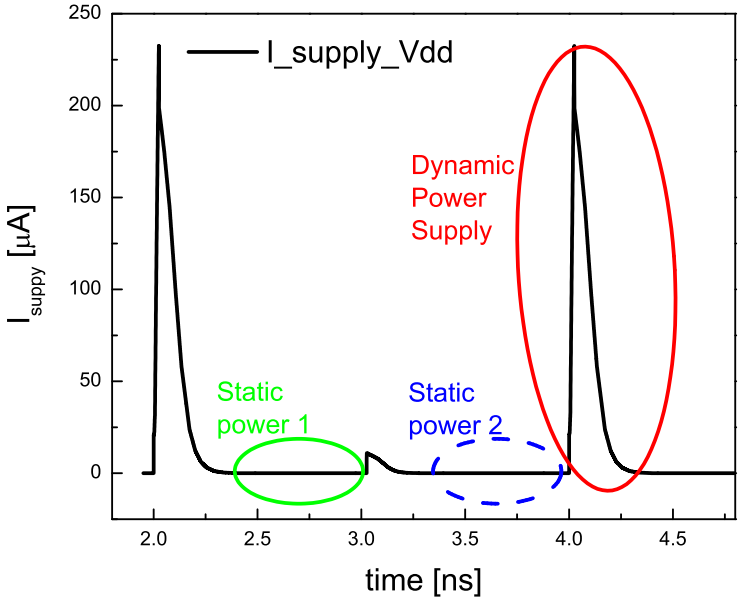


Figure 3.17: Current supplied through voltage source V_{dd}.

Fig. 3.17 shows one zone of dynamic power supply and two zones of static power. Table 3.8 shows the dynamic power supplied by voltage source of V_{dd}. The static part is in Table 3.9.

Simulation results in Table 3.8 and 3.9 show that EGL can not work well and may have serious error for such simulation. Around 30% of errors in static power supply are observed by both ON EGL and OFF EGL.

Table 3.8: Dynamic power supplied by voltage source.

Dynamic	Mean Power		Max Power	
	(μW)	diff.	(μW)	diff.
Post-litho Model	80	-	279	-
EGL ON	79.9	-0.13 %	273	-2.20 %
EGL OFF	80	-0.13 %	296	6.20 %
Post-litho Model w/ Variations	80.2	0.24 %	325	16.5 %

Table 3.9: Static power supplied by voltage source.

	Static 1		Static 2	
	(μW)	diff.	(μW)	diff.
Post-litho Model	36.8	-	53.7	-
EGL ON	37.4	-1.77 %	36.2	-32.63 %
EGL OFF	48.2	31.0 %	63.9	19.04 %
Post-litho Model w/ Variations	38.9	5.69 %	700	$12X$

Litho variations will have serious affect on the power consuming supplied by voltage distribution network, with 16.5% of increase in peak value of dynamic power supply, and 12 times of static power supply are induced by lithographic process variations. The explicit increase of static power supply is because the leakage of PMOS will exponentially increase with the decrease of gate length. And in static zone 2, V_{gs} of PMOS is zero, V_{ds} of PMOS is -1.2V, PMOS is off, and leakage current is the most sensitive to the gate length variations.

3.5.7 Inverter Chain Simulation

An inverter chain of 7 stages is simulated to test the affect and necessity of post-litho circuit simulation. The gate widths of PMOS and NMOS are $400nm$ and $200nm$, while the gate length is $65nm$, and the supply voltage is $1.2V$. The load capacitors are of $6fF$, as shown in Fig. 3.18. The input slew is $60ps$.

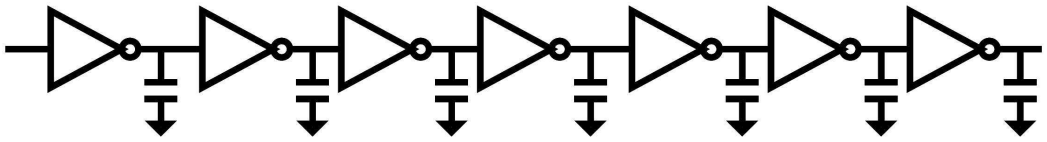


Figure 3.18: Inverter Chain.

The simulation results of output signal on stage 1 and 7 are shown in Fig. 3.19. A comparison of the delay/slew differences between EGL and our post-litho simulations on stage 1 and 7 shows that the errors of EGL method can be propagated and accumulated. The difference of delay and slew on stage 1 output between EGL and our proposed post-litho simulation is -5.36% and -3.92%, while the delay and slew difference on stage 7 output is -7.93% and -5.78% both of which are bigger than the errors on stage 1.

3.6 Summary

For the post-layout/post-lithography simulation, a new non-rectangular device model and corresponding circuit simulation methodology are proposed by using the drain-source current modification. The effect of non-rectangular

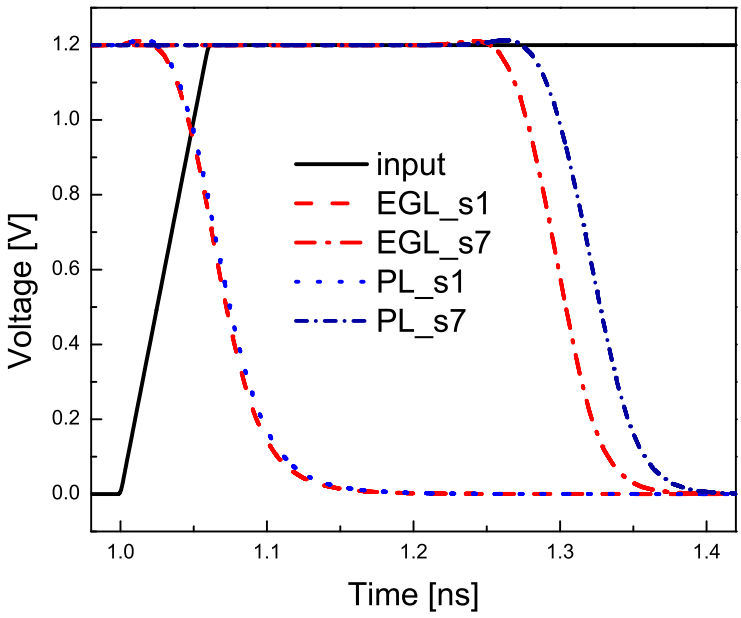


Figure 3.19: Delay comparison in inverter chain.
The output signals compare the *EGL* method and the proposed post-litho circuit simulation method (*PL*) on stage 1 (*s1*) as well as stage 7 (*s7*). The delay errors of EGL on stage 1 to 7 increase from 5% to 8%, and slew errors increase from 4% to 6%.

gate shape during parameter extraction of the device model is also considered for the first time. As far as we know, this is the most accurate methodology for post-litho analysis, including timing, leakage and transient simulation while it is well compatible to traditional modeling and simulation flow. The proposed model is validated and compared to the existing equivalent gate length (EGL) methodology. My simulation results show that EGL methods may lead to serious errors on both timing estimation (2%) and leakage/power estimation (up to 1.5X). My non-rectangular model provides a unified and accurate extraction, characterization, and simulation flow for both timing and power. Given

that nanoscale devices are becoming more and more non-rectangular, it is expected the unified post-litho non-rectangular model will be very useful for accurate timing and power analysis in future nanometer designs.

Chapter 4

Nanoscale Interconnect with Scattering Effect

For current very large scale integration (VLSI) circuit, interconnect has become one of the dominant factors in the overall circuit performance. With the continuous scaling down of feature size, the impact of interconnect delay is reduced in a much slower way than the intrinsic gate delay [3, 78] as shown in Fig. 4.1. Thus, different interconnect delay models are developed, such as Elmore delay [79, 80], AWE model [81, 82], and others [83].

To reduce the interconnect delay, new processing techniques like copper wiring have been adopted [84] and more advanced material like carbon nanotube has been investigated [85–87]. In EDA area, techniques such as buffer insertion, wire sizing and so on. have become indispensable. As a popular method to reduce the interconnect delay, wire sizing as well as shaping has been widely investigated, e.g., the wire sizing and shaping without fringing capacitance [88, 89], with fringing capacitance [90, 91], under transmission line model [92], or with single-width sizing (1-WS) or two-width sizing (2-WS) [93]. Wire sizing for multi terminal nets has also been extensively studied [94–96]. However, all these previous wire sizing algorithms are also based on the constant resistivity model discovered by Ohm in 1827 [97] without considering the

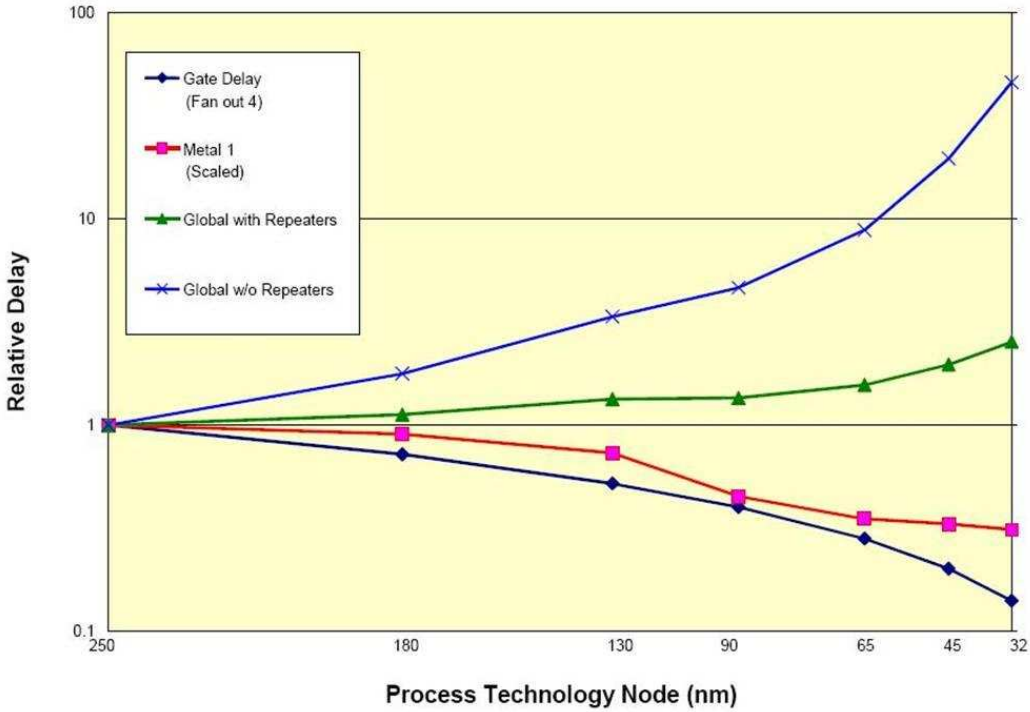


Figure 4.1: Delay for Metal 1 and Global Wiring versus Feature Size[3].

scattering effects in nanometer scale VLSI design.

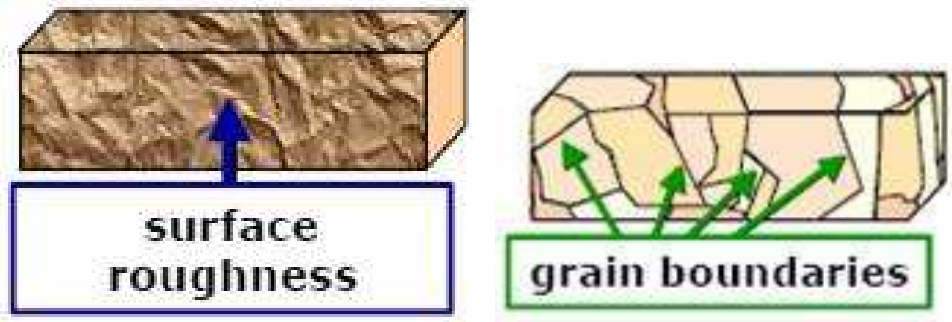
For nanometer scale interconnect, the scattering effect will first become prominent due to scaling. It will increase the effective resistivity and thus affect interconnection delay significantly. Until 2005, existing works on scattering effect are mostly performed using very complicated physics-based models, yet the scattering effect on nanoscale VLSI interconnect and optimization has not been studied. In my paper [98], I first presented a simple, closed-form scattering effect resistivity model based on extensive empirical studies with measurement data. When the proposed scattering model is ap-

plied to revisit several classic wire sizing/shaping problems, I found that not only the mathematical functions but also the delay values as well as optimization strategy would be changed dramatically. My experimental results showed that if the scattering effect is ignored or characterized inaccurately beyond 65nm, the resulting interconnect optimization might be away from the real optimal solution, for example, missing the scattering effect would lead up to 70% underestimation of the delay, or $20X$ over sizing. I also obtained a new closed-form wire sizing function taking into consideration of scattering effects.

4.1 Physics of Scattering Effect

As the feature size continues to shrink, the lateral dimension of conductors will be approaching to the mesoscopic regime in which the diameter of the wire is in the range of or smaller than the mean free path of the electrons (λ, about $40nm$ for copper at room temperature), and the electrical resistivity of metallic conductors is increased compared to the resistivity of bulky metal. The earliest work on this phenomenon dates back to 1938, when an resistivity expression of metal thin films (1-dimensional) was derived by Fuchs [99]. After that, it was extended to 2-dimensional by the *FS* model [100] for thin/narrow wires. Basically, the *FS* model accounts for the surface scattering. Later on, the *MS* model [101] was developed to incorporate the grain boundary scattering, which also increases the wire resistivity. Both surface scattering and grain boundary scattering effects are illuminated in Fig. 4.2.

Very complicated quantum mechanical effects can be applied to obtain



(a) Surface Roughness. (b) Grain Boundaries.

Figure 4.2: Two key mechanisms of scattering effect[4].

the empirical parameters in *FS* or *MS* model [102, 103]. For example, the surface scattering is modeled as 4.1 in [104].

$$\begin{aligned}\frac{\rho_0}{\rho} = & \frac{3}{4\pi hw}\int_{-h/2}^{h/2} dy \int_{-w/2}^{w/2} dx \int_{-\pi+\arctan(w/h)}^{\arctan(-w/h)} d\varphi \cdot f_{grain}(w,\varphi) \\ & + \frac{3}{4\pi hw}\int_{-h/2}^{h/2} dy \int_{-w/2}^{w/2} dx \int_{\arctan(-w/h)}^{\arctan(w/h)} d\varphi \cdot f_{grain}(h,\varphi)\end{aligned} \tag{4.1}$$

where,

$$f_{grain}(s,\phi) = \int_0^{\pi} \frac{1-\exp\{-s/[2\lambda\cos(\theta)\cos(\varphi)]\}}{1-p\cdot\exp\{-s/[2\lambda\cos(\theta)\cos(\varphi)]\}} \sin(\theta)\cos^2(\theta)\, d\theta$$

p is the proportion of electrons specularly reflected from the surface. w is the width and h is the height of the wire. ρ_0 is the bulk resistivity value and ρ is the resistivity of nanometer scale wire.

While *MS* and *FS* models have been tested with measurement data for polycrystalline nanowires [104–106], very few experimental results on copper

(Cu) film or interconnect have been reported until recently, e.g., the size effect of copper thin film was studied in [6], and the resistivity of copper wires with width under $50nm$ was reported in [5, 107–109]. The key observation is that the resistivity of copper wires will increase significantly as the wires width decreases [5, 6, 107–110]. While exploratory structures such as carbon nanotubes are being studied as possible substitution of copper interconnect [85, 111, 112], at least by 22nm technology, it is not likely that copper will be replaced by carbon nanotubes or other materials [85]. There are also efforts in manufacturing improvement to partially reduce the scattering effect of copper wires [113, 114], but even so the scattering effect can no longer be ignored as the technology continues to scale down.

4.2 Modeling of Scattering Effect for Interconnect Delay

For deep sub micrometer (DSM) scale, different interconnect delay models are developed, such as Elmore delay [79, 80], AWE model [81, 82], and others [83]. However, most of delay modeling works were developed twenty years ago and no quantum effect or scattering effect is considered in previous models. When scaling down to nanometer scale, new interconnect delay models are necessary to be developed. However, the complicated resistivity models (such as surface scattering model in Eq. 4.1) are too expensive and difficult to be adopted in the practical and real interconnect delay calculation and wire sizing optimization. First of all, a simplified analytical model is presented as

following which would be desirable for VLSI physical design applications.

4.2.1 Simplified Closed-Form Model of Scattering Effect

Based on my empirical study on both MS and FS models and curve fitting, I obtained the following simple closed-form width-dependent resistivity model with scattering effect.

$$\rho(w) = \rho_B + \frac{K_\rho}{w} \tag{4.2}$$

Fig. 4.3 shows that the simple resistivity compact model fits well with the measured experimental data [5]. Fig. 4.4 is the fitting results based on the measurement data from [6]. And I also verified the simple compact model with the complicated model in ITRS 2004 [110].

In the rest of this dissertation, I will use the fitting parameters $\rho_B = 2.202\ \mu\Omega \cdot cm$, and $K_\rho = 1.030 \times 10^{-15}\ \Omega \cdot m^2$ based on [5]. Note that ρ_B is almost as same as ρ_0, as the scattering effect can be ignored when the dimension of Cu wire is large enough.

4.2.2 Preliminaries and Key Parameters

The driver of the interconnection is modeled as an effective resistance R_d connected to an ideal voltage source and a sink as a load capacitance C_L. The Elmore delay model [79, 115] the most efficient and widely used delay model, is applied to compute the interconnect delay in this dissertation. Generally for the step input, the Elmore delay of a sink in an RC tree gave an

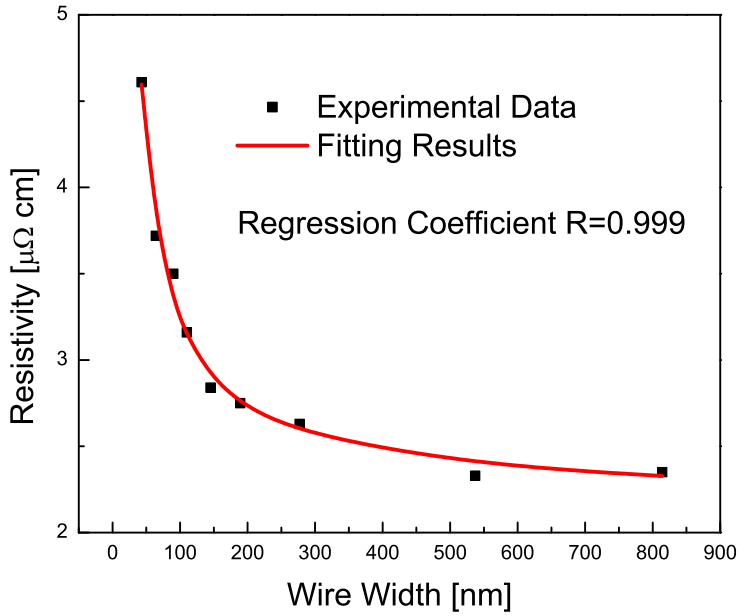


Figure 4.3: Resistivity fitting with scattering effect based on the experimental data[5].

absolute upper bound on the actual 50% delay of the sink [116]. The notations for the key interconnect and device parameters are listed in Table 4.1:

The values of basic parameters used are shown in Table 4.2. Note that these parameters are used mainly to illustrate the effect of scattering effect. The values are extracted according to [5, 93, 110, 117]. If necessary, more complete and specific parameters can be assigned.

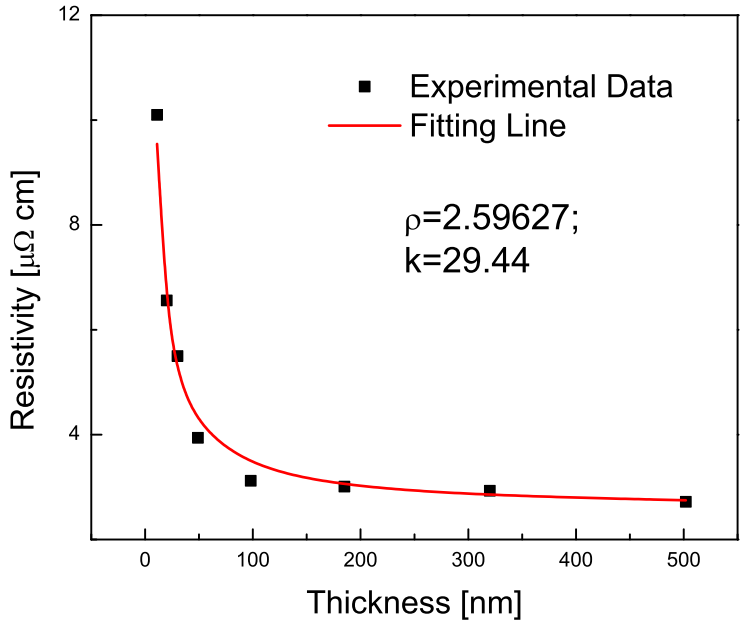


Figure 4.4: Resistivity fitting with scattering effect based on another experimental data[6].

4.2.3 Interconnect Delay with Scattering Effect

The delay of single width wire including scattering effect can be written as follows:

$$T_{1-WS} = R_d \cdot [c_a \cdot l \cdot w + c_f \cdot l + C_L] + \left[\frac{c_a \cdot l \cdot w}{2} + \frac{c_f \cdot l}{2} + C_L\right] \cdot \left[\rho_B + \frac{K_\rho}{w}\right] \cdot \frac{l}{w \cdot t} \tag{4.3}$$

where l is the wire length. In Fig. 4.5, the delay of the minimum width wires is calculated, and the ratio shows that the delay with scattering effect (T_{Wmin}) over delay without scattering effect ($T_{Win,NoS}$). Observe that the ratio is always greater than 1 and this implies that without considering scattering effect, the interconnect delay would be underestimated. Moreover, these ratios

Table 4.1: Variable notations of scattering effect.

Variable	Notations	Unit
$w_{\min}$	minimum wire width	nm
c_a	unit area capacitance	$fF/\mu m$
c_f	unit effective-fringing capacitance	$fF/\mu m$
c_g	input capacitance of a minimum device	fF
r_g	output resistance of a minimum device	$k\Omega$
ρ_0	resistivity of Cu, assume no scattering	$\mu\Omega \cdot cm$
$\rho_{\max}$	Max Cu resistivity of the minimum width	$\mu\Omega \cdot cm$
AR	aspect ratio	
t	metal thickness, $t = w_{\min} \cdot AR$	
R_d	driver resistance, $R_d = r_g/100$	
C_L	load capacitance, $C_L = 100 \cdot c_g$	

become larger with decreasing feature size. We can see that in nanoscale manufacturing, scattering effect must be considered. Otherwise, interconnect delay may be underestimated significantly, and cause serious timing error.

Table 4.2: Basic parameters of scattering effect.

Year	$w_{\min}$	c_g	r_g	c_a	c_f	ρ_0	$\rho_{\max}$	AR
2004	90	0.0625	24.09	0.056	0.04	2.2	3.35	1.7
2007	65	0.0573	22.75	0.056	0.04	2.2	3.79	1.8
2010	45	0.0375	24.82	0.056	0.04	2.2	4.49	1.8
2013	32	0.0246	27.07	0.056	0.04	2.2	5.42	1.9
2016	22	0.0161	29.53	0.056	0.04	2.2	6.88	2

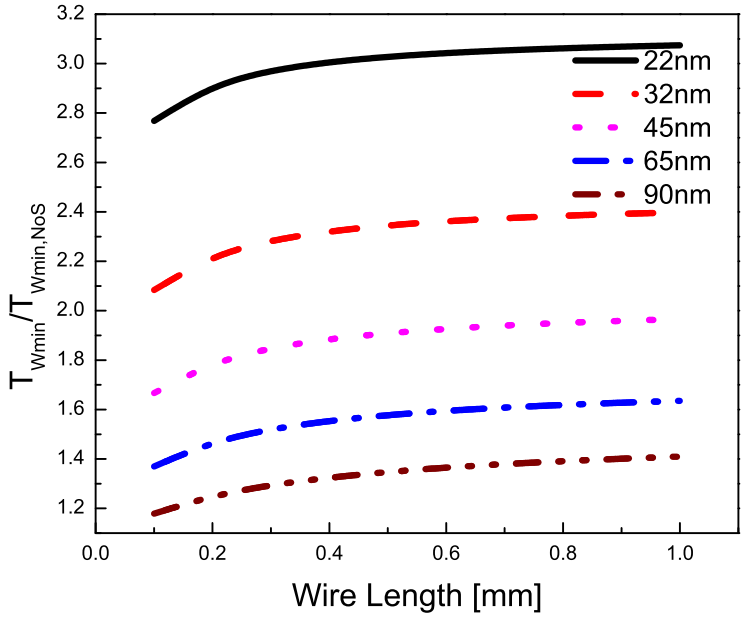


Figure 4.5: Normalized delay of different wire lengths under minimum wire width.
The normalized delay is the ratio of delay with scattering effect (T_{Wmin}) to delay without scattering effect ($T_{Win,NoS}$). Observe that the ratio is always greater than 1 and it worsens with decreasing feature size.

4.3 Wire Sizing with Scattering Effect

In this part, a new wire sizing formula based on the new width-dependent resistivity model with scattering effect Eq. 4.2 will be presented. My experimental results show that scattering effect will have major effect on wire sizing in nanoscale interconnections.

4.3.1 Efficiency of Wire Sizing

In the nanoscale interconnection, scattering effect induces bigger resistivity, which in turn causes bigger delay. Thus, wider wires become more

effective to compensate the more rapid increase of narrow wires resistance with scattering effect as wire sizing would give additional benefit on reducing resistivity). In other words, wire sizing considering scattering effect becomes more efficient and more necessary. The efficiency of wire sizing can be defined as gradient of $\partial T/\partial w$, the sensitivity of delay reduction due to wire sizing. Eq. 4.4 is the traditional wire sizing efficiency with a constant resistivity (maximum resistance of minimum wire width). And Eq. 4.5 is the wire sizing efficiency considering width-dependent scattering effect.

$$\left.\frac{\partial T(w,l)}{\partial w}\right|_{NoScattering} = R_d \cdot c_a \cdot l - \frac{l}{w^2} \cdot \left[\frac{c_f \rho_0 l}{2t} + \frac{C_L \rho_0}{t}\right] \tag{4.4}$$

$$\begin{aligned}\left.\frac{\partial T(w,l)}{\partial w}\right|_{Scattering} &= R_d \cdot c_a \cdot l - \frac{l}{w^2} \cdot \left[\frac{c_f \rho_B l}{2t} + \frac{C_L \rho_B}{t}\right] \\ &- K_\rho \cdot \frac{l^2}{w^2} \cdot \frac{c_a}{2t} - K_\rho \cdot \frac{l}{w^3} \cdot \left[\frac{c_f l}{t} + \frac{2C_L}{t}\right] \\ &\approx \left.\frac{\partial T(w,l)}{\partial w}\right|_{NoScattering} \\ &- K_\rho \cdot \frac{l^2}{w^2} \cdot \frac{c_a}{2t} - K_\rho \cdot \frac{l}{w^3} \cdot \left[\frac{c_f l}{t} + \frac{2C_L}{t}\right]\end{aligned} \tag{4.5}$$

where, $\partial T\,(w,l)/\partial w|_{NoScattering}$ is the gradient of delay over wire size without considering scattering effect, whose negative value would show the benefit of delay reduction by wire sizing. And $\partial T\,(w,l)/\partial w|_{Scattering}$ is the gradient with the consideration of scattering effect at nanometer scale interconnect. However, as a fundamental physical effect, the resistivity of nanometer scale wire always increased because of the surface or grain boundary scattering.

The most promising conclusion after comparing Eq. 4.4 and Eq. 4.5 is the additional negative items in Eq. 4.5. In Eq. 4.4, the first item $R_d c_a l$ is

always positive and constant which would increase the delay when sizing up the wire. However, the seconde item $-\frac{l}{w^2} \cdot \left[\frac{c_f \rho_0 l}{2t} + \frac{C_L \rho_0}{t}\right]$ is always negative which would help to reduce the delay by wire sizing. Unfortunately, its absolute value would reduce when wire width w is large. This means we can not unboundedly reduce delay by wire sizing (even under theoretical conditions) as the marginal benefit would be reduced. Eq. 4.5 shows that the additional items would always contribute negative value to the gradient and help to reduce the delay. This is the key reason why we introduce wire sizing/shaping with scattering effect into nanometer scale interconnect delay optimization. Fig. 4.6 shows the increased efficiency of wire sizing after considering scattering effect. Note that the wire sizing efficiency ratio of considering scattering over non-scattering is always bigger than 1. And for more advanced technology nodes, it might be possible to receive additional benefit of delay reduction by wire sizing.

4.3.2 Single Width Wire Sizing

In previous wire sizing and planning works [93], resistivity is assumed to be constant, and the optimal single width sizing (1-WS) can be calculated by:

$$w_{optimal,NoScattering} = \sqrt{\frac{c_f \cdot l + 2 \cdot C_L}{2 \cdot R_d \cdot c_a \cdot t} \cdot \rho} \tag{4.6}$$

In nanoscale interconnection, the scattering effect is prominent. In order to calculate nanoscale interconnection delay, the resistivity of bulky Cu should be replaced by a bigger value to avoid underestimation of delay. But if we put Eq. 4.6 into the use of wire sizing of nanoscale interconnection only by

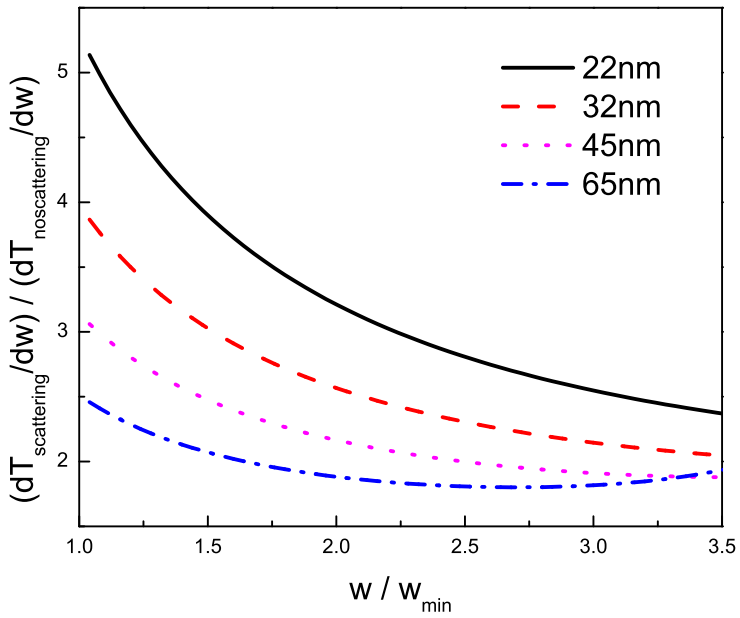


Figure 4.6: Compare the efficiency of wire sizing based on scattering and non-scattering.
The efficiency is defined as the gradient of delay over width. In this case, wire length is $10\mu m$. Observe that because of scattering effect, wire sizing becomes more efficient to reduce interconnection delay.

replacing old resistivity with a bigger value (such as ρ_{MAX}, which is the resistivity of the minimum width wire of certain specific technology node) and regarding this value as a constant, the optimal width of nanoscale Cu wire will be overestimated for the same reason that the resistance will decrease faster during wire width widening because scattering effect will be eliminated.

According to Eq. 4.3, to minimize the interconnection delay, $\partial T/\partial w = 0$. It becomes a cubic equation of wire width. Calculation shows that, from 130nm to 18 nm technology, only one real solution is bigger than the minimum

wire width. The analytical function is shown as below:

$$w_{optimal,scattering} = 2\sqrt{\frac{a_1}{3}}\cos\left(\frac{\vartheta}{3}\right) \geq w_{\min} \tag{4.7}$$

where,

$$\vartheta = cos^{-1}\left(\frac{K_\rho\left(c_f l + 2C_L\right)\sqrt{54R_d c_a t}}{\left(c_a K_\rho l + c_f \rho_B l + 2C_L\rho_B\right)^{3/2}}\right)$$

$$a_1 = \frac{c_a K_\rho l + c_f \rho_B l + 2C_L\rho_B}{2 \cdot R_d \cdot c_a \cdot t}$$

In Fig. 4.7, the optimal wire widths based on Eq. 4.7 and Eq. 4.6 are compared. As the $w_{optimal,no-scattering} - w_{optimal,scattering}$ could be as large as 10 times of minimum wire width and the ratios are always positive, it could be observed that the optimal wire width calculated based on Eq. 4.6 would be seriously overestimated as no scattering effect is considered in Eq. 4.6. This would cause excessive area waste and routability problem.

On the other hand, if we use a smaller constant resistivity ρ_0 for single width sizing (1-WS), the resulting so called optimal width will be less, and such optimized delay can be far away from the real optimal obtained based on the width-dependent resistivity caused by the scattering effect. Using Eq. 4.7, the optimal wire width can be calculated, and the result of wire sizing is shown in Fig. 4.8. The delay can be reduced by $50 - 90\%$ when using 1-WS optimization, compared to the minimum width.

Thus, it is very important to consider accurate scattering effect during the interconnect delay optimization and interconnect planning. Scattering effect must be considered with new equations like Eq. 4.7, or in the numerical

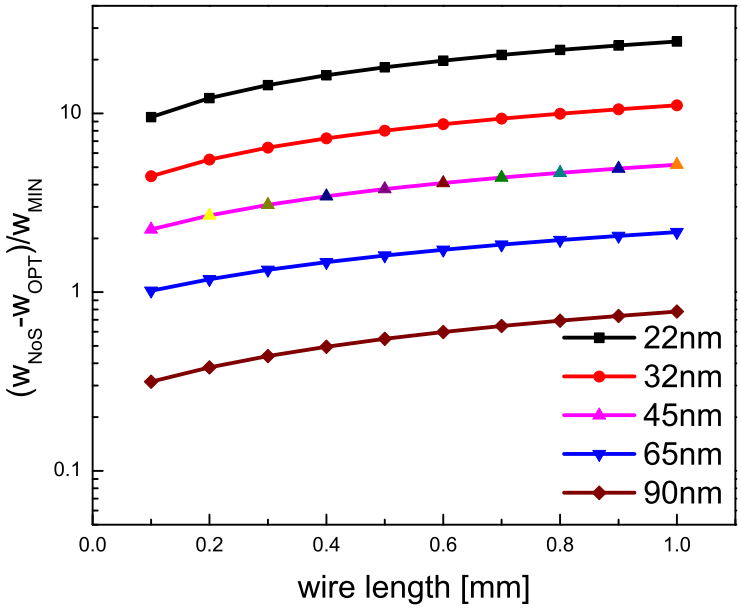


Figure 4.7: Comparison of optimal wire sizing.
The width difference is normalized by minimum width. Observe that it is always bigger than 0 and it worsens to more than $10X$ after $22nm$ node.

computation. And based on the width dependent resistivity model Eq. 4.2, it is not hard to extend the scattering effect to multi discrete wire sizing [93].

4.4 Wire Shaping with Scattering Effect

Even in paper [93], the authors found that under certain parameter settings, the optimal solution of two width sizing (2-WS) would be close to that of multi width sizing. However, this conclusion is based on the numerical calculation and specific parameter values. As a academic research, it is quite interesting and always critical to figure out what is the optimization limitation of wire sizing/shaping. Based on the understanding the theoretical limitation,

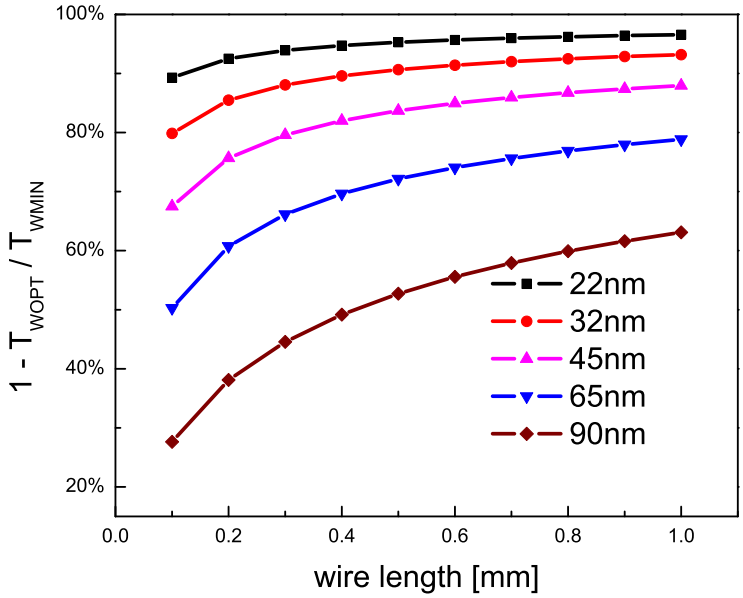


Figure 4.8: Delay reduction by wire sizing in nanoscale.
Observe that because of scattering effect, interconnect delay can be efficiently reduced by wire sizing.

it would be valuable to help the designers designers decide how much effort is required to afford the wire sizing optimization. In this section, wire shaping with scattering effect will be discussed.

Wire shaping (or continuous wire sizing) are discussed in previous papers without fringing capacitance [88, 89], or with fringing capacitance [90, 91] for deep sub-micrometer scale circuit design. As scattering effect was not considered, above works would lead to wrong solution for nanometer scale circuit design. Thus, I extended the work [88] with fringing capacitance to nanometer scale with scattering effect.

4.4.1 Euler's Differential Equation (math background)

To find the optimal solution of wire shaping is not a straightforward math issue. For example, the optimal wire shaping functions in [89, 91] are not close-form. In my approach, Euler's Differential Equation [118] is adopted. A quick review is listed as below.

Lemma 4.4.1. *If $u(x)$ is a function which can produce minimum I,*

$$I = \int_{x_0}^{x_1} F\left(x, u\left(x\right), u'\left(x\right)\right) dx$$

$u(x)$ should satisfy Euler's Differential Equation:

$$F_u\left(x, u\left(x\right), u'\left(x\right)\right) = \frac{d}{dx} F_{u'}\left(x, u\left(x\right), u'\left(x\right)\right)$$

Proof. See [118]. □

4.4.2 Wire Shaping to Minimize Elmore Delay

The wire shape can be defined as $f(x)$, and $x = 0$ is the terminal to contact driver R_d, the terminal to contact load C_l is $x = L$. L is the wire length, which is a constant in the wire shaping problem. As the scattering effect is considered, the function definition for Euler's equation is different from the previous work [90].

The wire shaping problem can be modeled similar to Lemma 4.4.1. The

delay to be optimized could be expressed as Eq. 4.8.

$$\begin{aligned} T &= \int_0^L \left[R_d \cdot C_L + c_a \cdot f(x) \cdot dx + c_f \cdot dx\right] \\ &+ \int_0^L \left[C_L + \int_0^x \left(c_a f(l) + c_f\right) dl\right] \left[\frac{\rho_B}{f(x)} + \frac{K_\rho}{f^2(x)}\right] \frac{dx}{t} \\ &= R_d \cdot C_L + \int_0^L F(x, u, u') \cdot dx \end{aligned} \tag{4.8}$$

where, $u(x) = \int_0^x f(l) \cdot dl$, and $u'(x) = f(x)$. $f(x)$ is the optimal wire shape that we are solving. To minimize the delay in Eq. 4.8, according to Lemma 4.4.1, Euler's equation can be rewritten as:

$$\begin{aligned} &c_a (u')^2 \cdot [2\rho_B \cdot u' + 3K_\rho] + c_f \cdot u' [\rho_B \cdot u' + 2K_\rho] \\ &= [C_L + c_a \cdot u + c_f \cdot x] [2\rho_B \cdot u' + 6K_\rho] \cdot u'' \end{aligned} \tag{4.9}$$

To apply the Eq. 4.9 practical for wire shaping application, polynomial approximation is used to express $u(x)$.

$$u(x) = \sum_{n=0}^{\infty} a_n \cdot x^n \tag{4.10}$$

where, a_0 is some constant; a_1 is the minimum wire width decided by the technology, and also the shape function $f(x)$ at location of terminal contacting driver R_d where $x = 0$ as:

$$\begin{aligned} f(x) &= u'(x) = \sum_{n=1}^{\infty} n a_n \cdot x^{n-1} \\ f(0) &= a_1 \end{aligned}$$

For the rest of a_n $(n = 2, 3, ...)$, they could be calculated by Eq. 4.9 one by one. For example,

$$a_2 = \frac{c_a {a_1}^2 (2\rho_B a_1 + 3K_\rho) + c_f a_1 (\rho_B a_1 + 2K_\rho)}{4C_L (\rho_B a_1 + 3K_\rho)}$$

4.4.3 Approximate Model of Wire Shaping

The continuous wire sizing/shaping function $f(x)$ can be approximated by $g_n(x)$ with different orders of accuracy.

$$\begin{array}{rcl} g_1(x) & = & a_1; \\ g_2(x) & = & a_1 + (a_2 x + a_3 x^2)/2; \\ & \cdots & \\ g_n(x) & = & \sum_{i=1}^{2n-3} a_i x^{i-1} + \frac{1}{2}\left(a_{2i-2} x^{2i-3} + a_{2i-1} x^{2i-2}\right) \end{array} \tag{4.11}$$

Fig. 4.9 shows the difference from $g_1(x)$ to $g_5(x)$ of $45nm$ technology. In most cases, the approximation using $g_3(x)$ is good enough to get the optimal wire shaping.

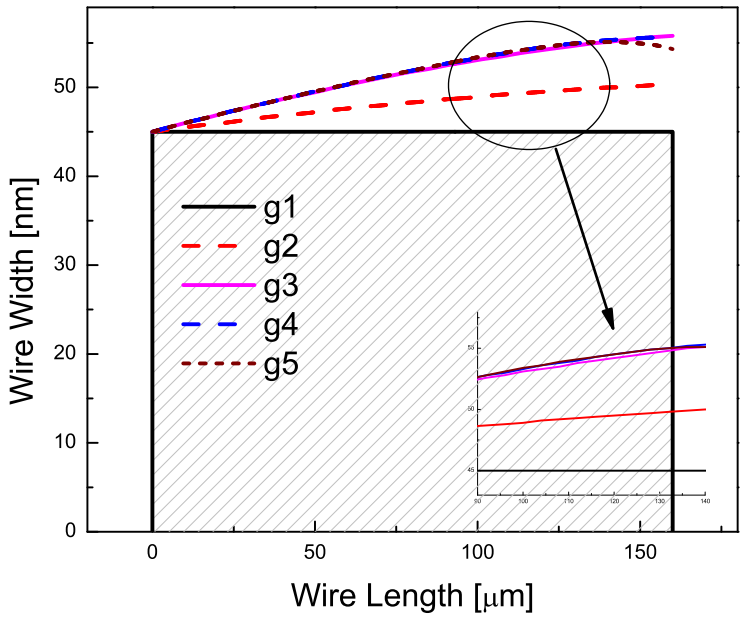


Figure 4.9: Wire shaping with different orders of polynomial approximation.

It should be noted that $g_n(x)$ is different from [90], e.g., $g_3(x)$ is a

quintic function while in [90] it is a cubic function. The scattering effect also increases the complexity of the approximated solutions.

Fig. 4.10 gives the delay comparison of different polynomial approximation orders in the wire shaping function. In this case, it can succeed to reduce the delay by 18% when using closed form $f(x) = g_3(x)$ of wire shaping rather than using minimum width $f(x) = g_1(x) = a1$. From Fig. 4.9, the area of wire only increase by 16% for wire shaping. And closed form $f(x) = g_3(x)$ is accurate enough for wire shaping because the deference between this closed form function and more accurate ones is no more than 1%.

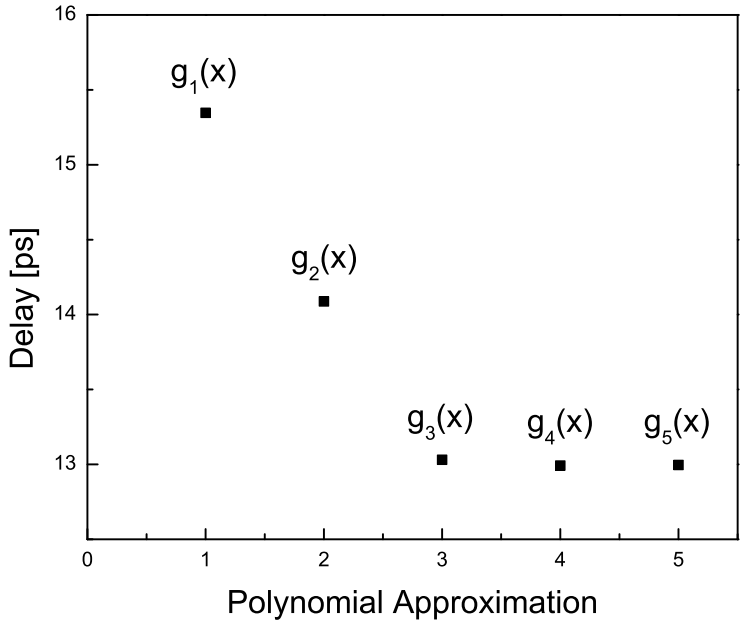


Figure 4.10: Delay estimation of wire shaping of $45nm$ node.
Comparison of different orders of polynomial approximation. Observe that the difference among g_3, g_4 and g_5 is less than 0.5%.

4.5 Summary

As interconnect delay is playing a more critical role in IC design, in this section modeling and optimization of nanometer scale interconnect have been studied. As an emerging topic for nanoscale interconnection, the scattering effect is discussed from different aspects. A simple and efficient width-dependent resistivity model due to scattering effect is firstly obtained and verified based on extensive empirical studies on measurement data. Then it is applied to the classic wire sizing problems and show the importance of considering scattering effects for future interconnect delay estimations and optimizations.

To my best knowledge, this is the first work that addresses scattering effects on the nanoscale interconnect sizing. This work would be helpful to reduce the gap between layout stage estimation/optimization and silicon data. It could be used as guideline to design the modeling and optimization strategy of nanometer scale IC interconnect.

Chapter 5

Latch Modeling for Statistical Timing Analysis

5.1 Introduction of Latch and Statistical Timing

Process variations pose the biggest challenge to technology scaling into nanometer regime as it is a major performance limiter. Statistical Static Timing Analysis (SSTA) has been proposed to perform full-chip analysis of timing under process variations and has been the subject of intense research recently [119–125].

In SSTA, the gate delays in the cell library are modeled as a first order approximation [122] or second order approximation [123] of process variations. Based on these models, statistical timing analysis and optimization can be applied to the combinational logic [124]. To attain more accuracy SSTA is done considering the clock distribution network [125]. By these approaches one can predict both the data signal's statistical distribution at the end of each combinational logic chain and the clock distribution at each clock network terminal. However, so far there is no work accurate enough to combine the signal distribution from both networks and predict final signal distribution of the whole system. The major reason is that there are no accurate delay models for the sequential logic such as Flip-flop and latch. Flip-flop and latch are the most

commonly used sequential elements whose purpose is synchronizing data signals. In other words, data signals get into a lock-step manner. These elements will add some delay to timing and thus decrease the system performance.

In this chapter, we focus on modeling latch accurately. This is because an edge-triggered flip-flop is functionally a back-to-back latch pair and also structurally made up of two latches in some applications [126]. Hence flip-flop models can be derived from accurate latch models.

A latch is a three-terminal element, with two inputs, data (D) and clock (clk) and one output (Q). The data must set up before the falling edge of the clock, which would give bigger and more flexible timing window for circuit design. For timing requirements, level sensitive latches are widely used in high performance ICs where timing analysis is more critical and challenging [127–129]. In these approaches the latch delay model is deterministically ignoring the fact that the input data signal and clock signal have statistical quantities. However, when a path is timing critical, the data would arrive very close to the falling edge of clock, and the mean value of t_{DC} (data-to-clock delay) might be close to the latch's setup time with very limited or negative slack left leading to the increase in the delay of data D to output Q (t_{DQ}). Moreover, with different slew distributions of data and clock, the t_{DQ} to t_{DC} function will be different. To keep things simple, traditional circuit design and timing analysis [130] choose a constant setup time. But this simplification leads to less accurate statistical timing analysis and even lesser flexibility in optimization [131].

In this chapter, we propose a new latch delay model for statistical timing analysis. Our latch model captures the effect of delay and slew variations of both input data and clock on latch delay. Based on this new latch delay model, one can combine the timing analysis of data signal network with clock distribution network to do SSTA in an accurate way. This provides a chance to do statistical timing of latch output Q by combinatorially analyzing statistical timing of both latch input data and clock as Fig. 5.1.

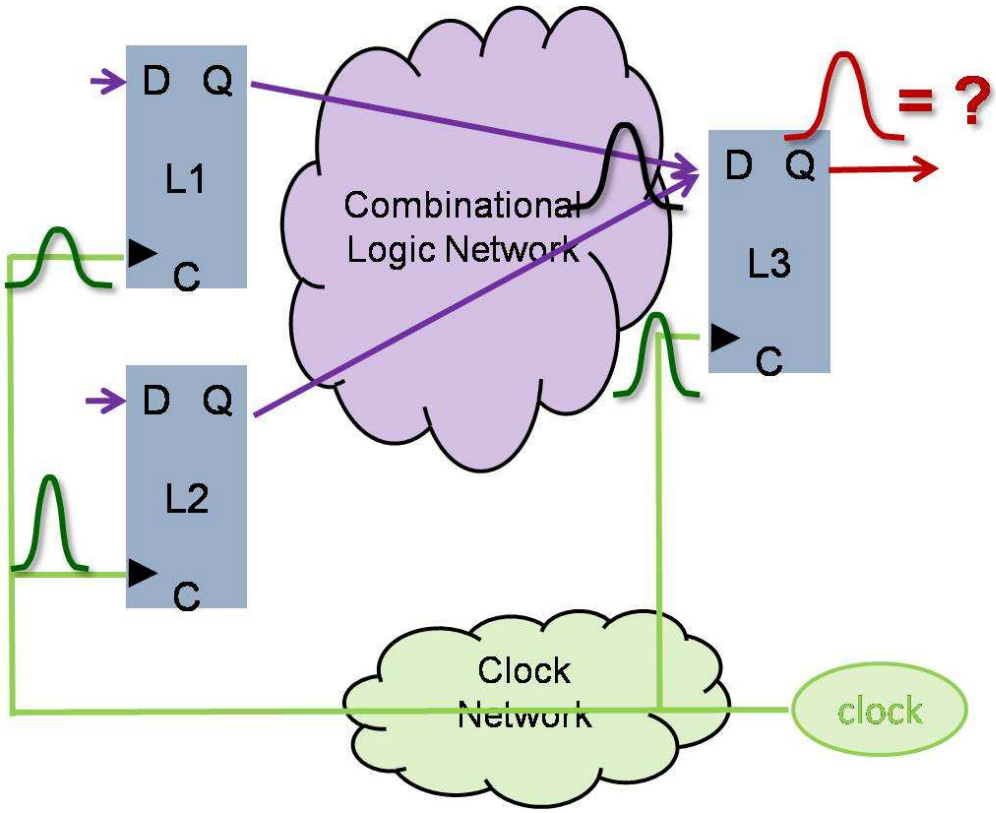


Figure 5.1: Combinatorially analyze statistical timing of both latch input data and clock.

The main contributions of this work include: a) a new latch timing

model considering both logic and clock signal variations; b) we integrate it into a unified SSTA framework. Our experimental results show that ignoring latch modeling may lead to large errors (e.g., 50% at PDF peak).

The rest of this chapter is organized as follows: in Section 5.2, general timing diagram and structure of transparent latch are reviewed, with traditional latch delay model. A new point of view for latch working mode based on a 3-D analysis is proposed in Section 5.3. Section 5.4 presents our new latch delay model and shows how to cover different data slew, clock slew and fanout from input signal's variations. Statistical timing analysis for latch is also discussed in Section 5.4, followed by experimental results in Section 5.5. Summary is presented in the last.

5.2 Latch Preliminaries

5.2.1 Timing Diagram of Latch

The timing diagram of latch is shown in Fig. 5.2,. Both setup and hold times of a latch are measured relative to the trailing edge of the clock. The data signal must be a constant in the timing window between the setup and hold time. This ensures that the data is sampled and latched correctly.

In addition to setup and hold times, two more delay quantities t_{CQ} and t_{DQ}, need to be defined. This is because of the following two scenarios: 1) Data is stable but the latch is closed due to the clock is low, and 2) Data gets stabile while the latch is open. In critical path analysis, when we assume that the data signals arrive quite close to the setup time while latch is open, t_{DQ} is

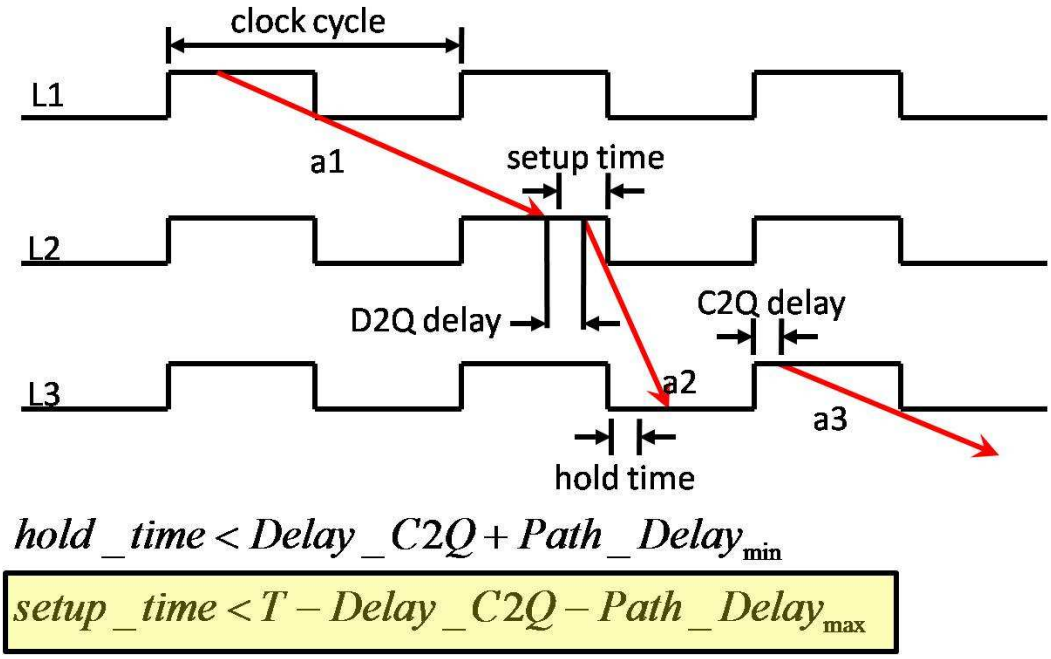


Figure 5.2: Timing diagram of latch.
The situation with the latch is different from flip-flop. Both setup and hold time of latch are measured relative to the tailing edge of the clock. The longest path "a1" must arrive at next latch "L2" before setup time and the shortest path "a2" must reach next latch "L3" after hold time.

the key delay to be analyzed. In this paper, we focus on modeling setup time and t_{DQ} accurately. The proposed method can be extended to hold time or other timing issues of latch.

5.2.2 Structure of Transparent Latch

Two of the most widely used latch structures are shown in Fig. 5.3 and 5.4. In the semi-custom datapath application, where the noise of the input signal can be well controlled, latch structure in Fig. 5.3 is preferable for

it is fast and compact. As mentioned in [132], Intel uses this as standard data path latch.

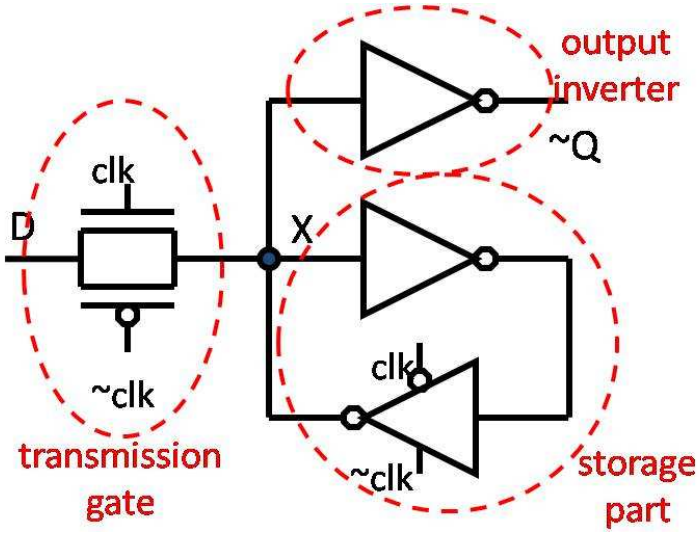


Figure 5.3: One of the most widely used latchs for its speed and compactness.

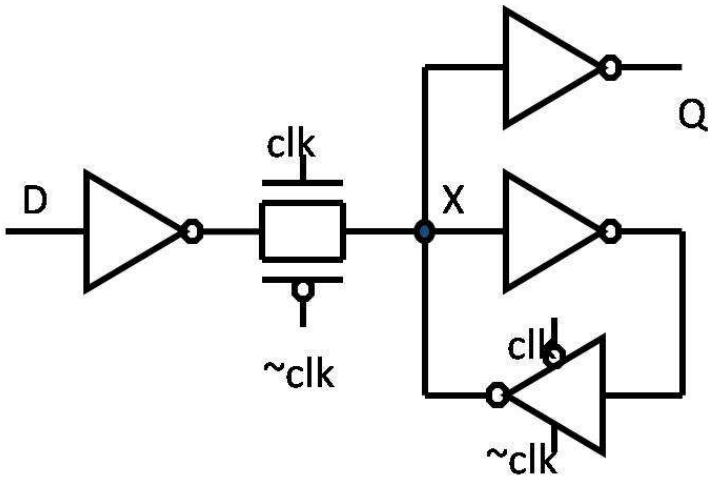


Figure 5.4: A widely used latch in standard cell applications.

With one addition inverter before the data input of latch in Fig. 5.3,

the latch structure in Fig. 5.4 is very robust transparent, which is widely used in standard cell applications [133]. Such a latch is recommended for all but the most performance-critical or area-critical design.

In this work, we focus on modeling the latch structure in Fig. 5.3 but our modeling is generically sufficient to be applied to the latch structure in Fig. 5.4 too.

The latch in Fig. 5.3 can be decomposed into 3 parts: the transmission gate, output inverter, and the storage part as marked in the Figure. Next, we will show that the traditional latch modeling focuses on the feedback mechanism of the storage part and models it as two inverters.

5.2.3 Traditional Timing Model of Latch

As shown in Fig.5.5, the traditional way of modeling latch focuses on the storage part of the latch [134], which is modeled as self-feedback system of two inverters.

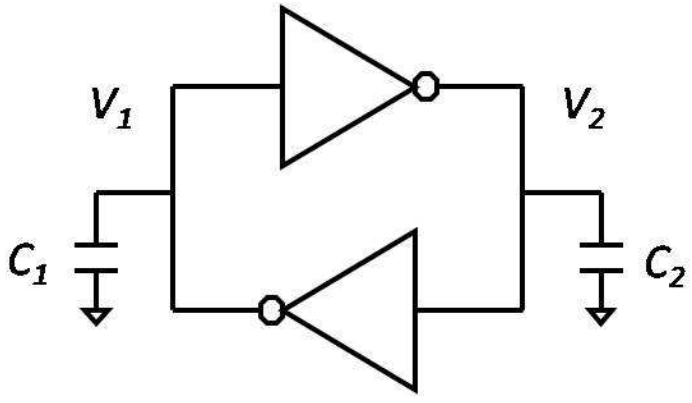


Figure 5.5: The storage part of a latch.

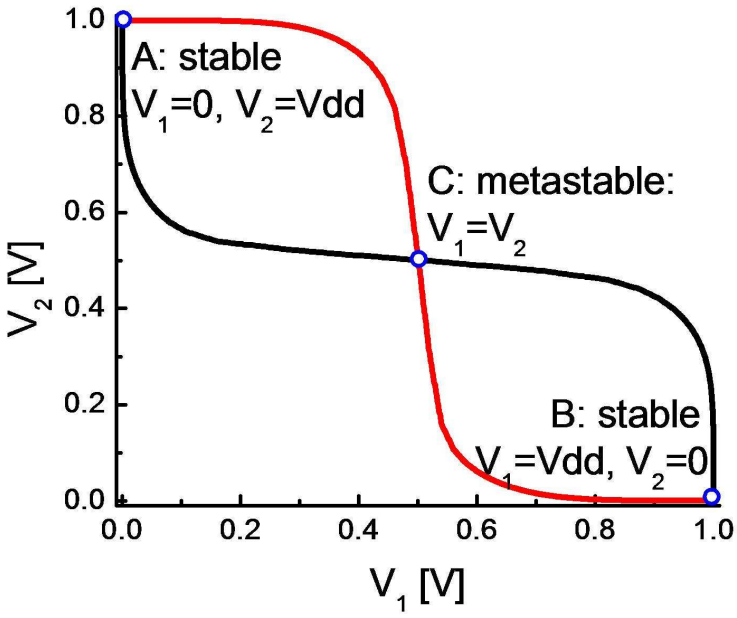


Figure 5.6: Butterfly curves of the static transfer characteristics.

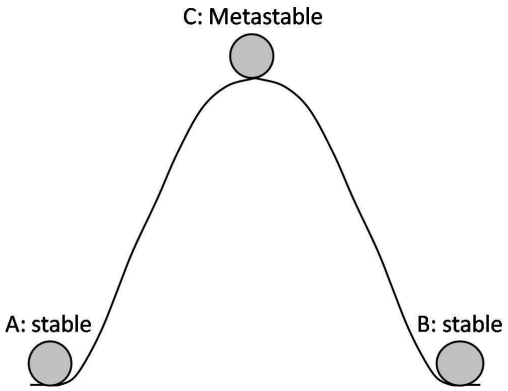


Figure 5.7: An analogy of a ball on a hill with one metastable state at the top of the hill and two stable states in the foothills.

Fig. 5.6 shows the famous butterfly curve of static states' transfer in Fig. 5.5. This feedback system has two stable states (point A and B) and one metastable state (point C) as shown in Fig. 5.7, based on which the *D2Q* delay

can be expressed as Eq. 5.1.

$$t_{DQ} = \tau_s \left[\ln \Delta V - \ln a\left(0\right)\right] \tag{5.1}$$

where t_{DQ} is the delay from input D to output Q, and $a\left(0\right)$ is a small signal offset from the original metastable point. ΔV is certain predefined constant voltage point to predict D-to-Q ($D2Q$) delay.

An additional assumption is that $a\left(0\right)$ is proportional to $(t_{DC} - t_m)$, where the input signal is a ramp that passes through the metastable state point at t_m. Thus, the $D2Q$ delay can be modeled as logarithmic function:

$$t_{DQ} = a - b \cdot \ln\left(t_{DC} + c\right) \tag{5.2}$$

5.2.4 Limitation of Traditional Model

To better understand the traditional model of the latch, several HSPICE simulations were run to get the delays of latch around setup time. We used PTM [69] for 65nm in our simulation. Then we fitted the resulting data according to Eq. 5.2 and the result is shown in Fig. 5.8.

In Fig. 5.8, the fanout of the latch is *4*, slew of the clock signal is $40ps$ and the slew of input data D is $80ps$. Black dots are HSPICE simulation results and the red line is the curve fitted based on traditional delay model in Eq. 5.2. Blue dash line is the input D-to-C ($D2C$) delay distribution that has positive slack as the mean value of $D2C$ delay bigger than setup time. The setup time is defined according to t_{DC} when t_{DQ} is 10% bigger than its minimum value.

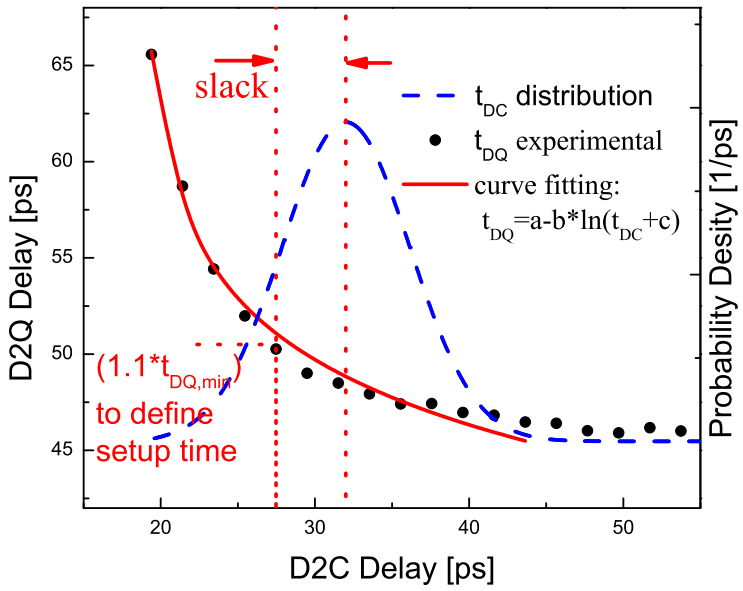


Figure 5.8: Limitation of the traditional latch model.
Traditional model is only accurate when D2C delay is much smaller than the setup time. However, for SSTA of critical paths, D2C delay is close to or bigger than the setup time.

From the figure, we can see that when *D2C* delay is around or bigger than setup time, the Eq. 5.2 is quite inaccurate. The fitting looks good only when *D2C* delay is much smaller than setup time. For statistical timing analysis of critical longest paths, as the mean of *D2C* delay is close to setup time and high percentage of *D2C* delay distribution will be around the setup time the latch, delay model of latch in Eq. 5.2 has difficulty to meet accuracy requirement of latches' statistical timing analysis.

Moreover, the model in Eq. 5.2 does not consider the effec of input data slew, clock slew or fanout. In fact, input data slew, clock slew and fanout, all of them could change the delay curves between t_{DQ} and t_{DC}.

5.3 A New 3D View of Latch Timing

5.3.1 State Transform in the Latch Storage Part

If the two inverters in the storage part of the latch are the same and driving strength of the PMOS and NMOS in each inverter are also identical, the potential of the storage part can be drawn as Fig. 5.9 and 5.10.

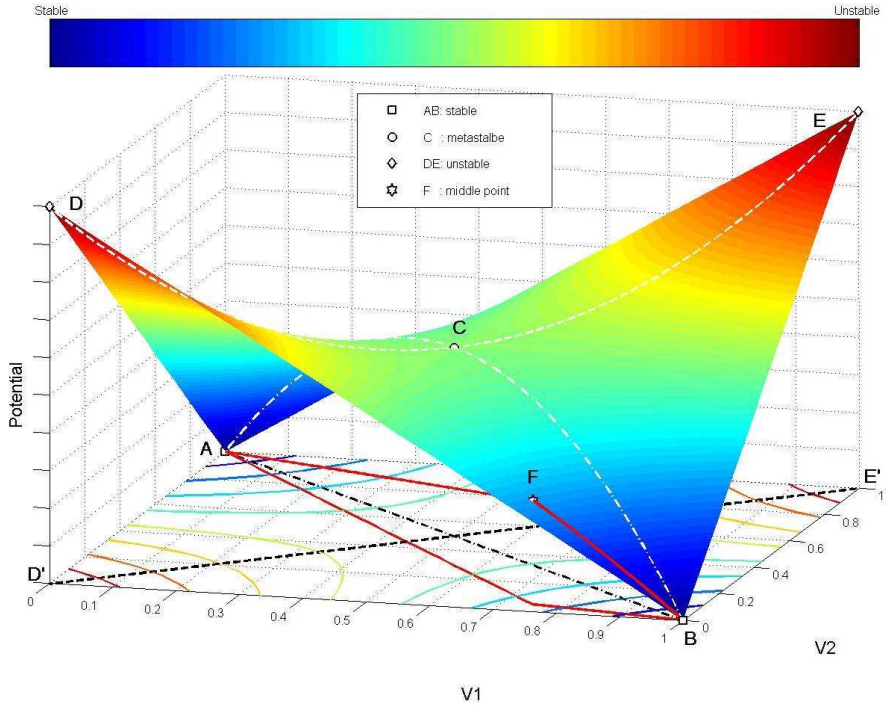


Figure 5.9: 3D potential figure with 2D projection.
Traditional latch delay function models the state transfer along ***A***-***C***-***B***, where ***A*** and ***B*** are two stable states and ***C*** is the only metastable point. However, we point out it is much more possible that the storage part of latch will be driven by the transmission gate directly from ***A*** to some middle point ***F*** far away from ***C***, and then slides from ***F*** to ***B***.

In Fig. 5.9, the 3D potential is drawn while X and Y axis are V_1 and V_2 respectively. 2D projection is also drawn in Fig. 5.9 and amplified in Fig. 5.10.

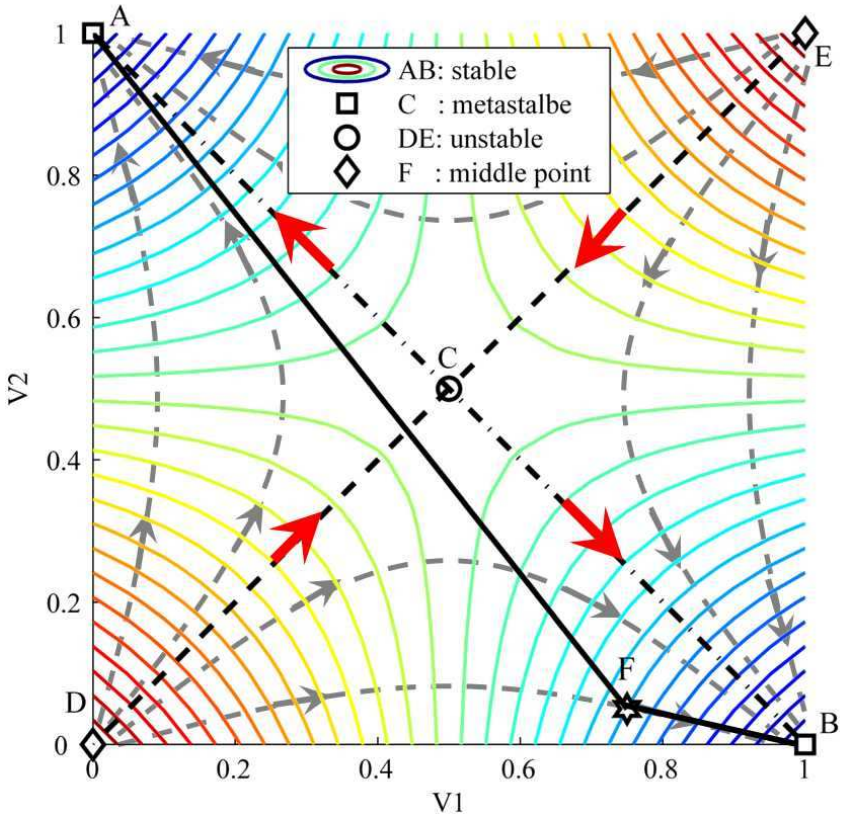


Figure 5.10: 2D square amplification of potential projection.

There are 5 special state points:

- ***A***: $(V_1 = 0,\ V_2 = V_{dd})$, stable;
- ***B***: $(V_1 = V_{dd},\ V_2 = 0)$, stable;
- ***C***: $(V_1 = V_2 = V_{dd}/2)$, metastable;
- ***D***: $(V_1 = V_2 = 0)$, unstable with highest potential;
- ***E***: $(V_1 = V_2 = V_{dd})$, unstable with highest potential.

D' and ***E'*** are the projections of ***D*** and ***E*** on 2D plane.

When the state of the storage part is at point ***A*** or ***B***, the state is stable at lowest potential. Point ***C*** is the only metastable state in the system.

Traditional latch model in Eq. 5.1 only covers the state transfer from one stable state through metastable point and to another stable point, which is the white dash dot line ***A***-***C***-***B*** in 3D potential of Fig. 5.9 (also the black dash dot lines on the 2D projection in Fig. 5.9 and Fig. 5.10).

On the project plane of square ***A***-***D***-***B***-***E*** in Fig. 5.10, there are much more state points than the points on line ***A***-***C***-***B***. The colored solid lines show the equipotential lines. The dash lines show that the state moving tracks if there is no external signal input. For example, if the state is at $\boldsymbol{D}(V_1 = V_2 = 0)$ or $\boldsymbol{E}(V_1 = V_2 = V_{dd})$, it will directly go to the metastable point ***C*** along the black line ***D***-***C*** or ***E***-***C*** with red arrows, and then through ***C*** go to stable states of ***A*** or ***B***. During this process ***D***-***C*** or ***E***-***C***, if there is any noise, the state transfer will follow along the grey dash lines in Fig. 5.10) and go to stable points ***A*** or ***B*** directly.

According to the above analysis, simplification in traditional latch model is rough and wrong for modeling the state transforming process, thus even using curve fitting, the model in Eq. 5.2 has difficulty to fit the simulation results around setup time.

5.3.2 Practical Latch Simulation

Fig. 5.11 shows the voltage changing of each node of the latch structure in Fig. 5.9 based on SPICE simulation. The voltage transfer of point X (see Fig. 5.9) can be divided into two parts. At first, V_1 is changed almost linearly until arriving certain middle point $\boldsymbol{F}$. After reaching $\boldsymbol{F}$, V_1 increases at a gradually slow speed. At the same time, V_2 is changing in a different way.

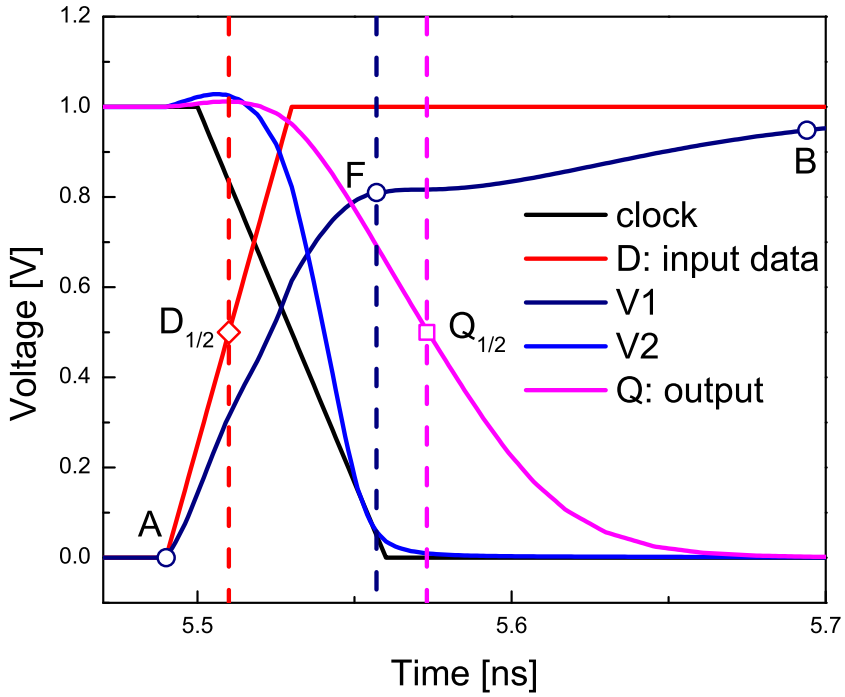


Figure 5.11: Voltage curves of each node in latch.
D2Q delay is made up of 2 parts:
1) from $\boldsymbol{D}_{1/2}$ to $\boldsymbol{F}$, which is driven by input data signal;
2) from $\boldsymbol{F}$ to $\boldsymbol{Q}_{1/2}$, which is a self-feedback process.

From the figure, we can see, as clock turns off the inverter from V2 to V1, V2 increases to its targeting final stable state in a faster speed than V1,

thus in Figure 5 (b), the position of F is lower than line C-B. If the input data signal is close to the setup of the latch, the state transfer of the latch storage part is in following ways:

1. Driven by the input data signal current through the transmission gate, the storage part of the latch is moved to state point of ***F***. During this process, the storage part will move from stable state ***A*** to ***F*** directly instead of through the metastable state of ***C***. This process is more likely to be linear rather than logarithmic.

2. Afterwards the clock turns on the inverter from V_2 to V_1, and the storage part turns to self-feedback and move from ***F*** to ***B*** at a gradually slow speed. The traditional latch modeling discussed in Eq. 5.2 focuses on this part and assumes state point ***F*** is on the state transforming track ***C*** to ***B***.

When both the delay and slew of the input data as well as clock signals are statistical, it will be too time consuming to run SPICE for each case. Based on the above visual understanding of latch timing, a new latch model for external variation is proposed in the next section.

5.4 A New Latch Model for t_{DQ} Delay

In this section, a new latch delay model will be proposed based on the 3 dimensional view of latch state transform. The new delay model will focus on the variable of the clock-to-input ($C2D$) delay. In this way, the clock

skew and statistical timing of combinational logic delay could be considered. Then the model will be extended by involving the clock slew and input data slew to input accuracy. Finally, the lithographic variations of non-rectangular gates of latch itself are also involved. The variations of clock and input data signals could be regarded as external and the lithography related variations from latch's own gates are internal. The proposed latch delay model could consider both external and internal variations finally.

5.4.1 Difficulty of Latch Modeling

As discussed in previous section, the latch state transfer from one stable to another stable state can be divided into two parts, ***A***-***F***: driven by input data signal in a close to linear function, and ***F***-***B***: more like a logarithmic or exponential function which is self-feedback process of storage part in latch.

However, it is very difficult to develop pure analytical function for latch modeling. As the gate length of device moves to 65nm and below, the device model becomes much more complicated, and the gates, such as inverters and transmission gate are far away from linear to input voltage. The 3D camber in Fig. 5.9 will be distorted and the metastable point will be moved away from position ***C***.

SRAM (which is similar to the storage part of latch) is modeled as dynamic system and gives an analytical function to predict critical time of noise [135]. However, the input signal's current waveform is complicated and can not be modeled as square wave. And the inverters in the practical latch

are skewed whose PMOS and NMOS have different driving strengths. As only a few special functions can be solved in dynamic system [136], all of these properties in the practical latch challenge the effort to deduct pure analytical function for latch modeling.

Thus in this work, instead of deducting pure physical analytical model, a semi-empirical function for latch modeling is proposed to cover all of the affects including not only *D2C* delay but also input data slew, clock slew and fanout.

5.4.2 Three Regions of $t_{DQ} - t_{DC}$

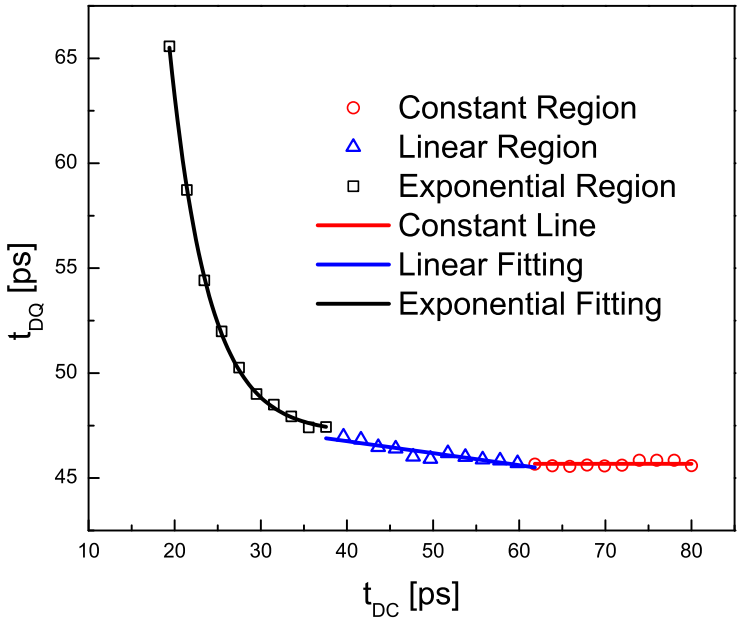


Figure 5.12: 3 regions of latch delay curve:
constant region (red line/round dots), linear region (blue line/triangle dots), and exponential decay region (black line/square dots).

In consequence the whole $t_{DQ}\,(t_{DC})$ can be divided into three regions (Fig. 5.12).

1. Constant region (red line/round dots). In this region the latch is absolutely transparent and *D2Q* delay is a constant. During this process, clock is on, and input data signal goes through latch directly to ***Q***.

2. Linear region (blue line/triangle dots). With the decrease of *D2C* delay, the transmission gate is open for quite long period, and the input data signal drives the storage part from stable state (such as ***A***) to certain middle point ***F*** which is close to another stable state (such as ***B***). In this process, the part directly from ***A*** to ***F*** dominates the *D2Q* delay.

3. Exponential decay region (black line/square dots). In this region, the process from ***F*** to ***B*** is dominant in the total *D2Q* delay.

5.4.3 Latch Delay Modeling Function

The proposed latch model is divided into two parts: when t_{DC} is big enough, t_{DQ} is constant; after t_{DC} gets smaller, the model is made up of two components: linear part and exponential decay part, shown as

$$t_{DQ} = \begin{cases} t_{DQ0} & t_{DC} \geq t_{DC0} \\ a \cdot \exp\left(-b \cdot t_{DC}\right) + c \cdot t_{DC} + d & t_{DC} < t_{DC0} \end{cases} \tag{5.3}$$

where

$$t_{DQ0} = a \cdot \exp\left(-b \cdot t_{DC0}\right) + c \cdot t_{DC0} + d$$

If the variations of data slew, clock slew and fanout are within a small range or a large approximation is acceptable during the statistical timing analysis, model in Eq. 5.3 can be simplified to some exponential decay function such as:

$$t_{DQ} = \begin{cases} t_{DQ0} & t_{DC} \geq t_{DC0} \\ a_1 \cdot \exp\left(-b_1 \cdot t_{DC}\right) + d_1 & t_{DC} < t_{DC0} \end{cases} \tag{5.4}$$

where

$$t_{DQ0} = a_1 \cdot \exp\left(-b_1 \cdot t_{DC0}\right) + d_1$$

Or even,

$$t_{DQ} = a_2 \cdot \exp\left(-b_2 \cdot t_{DC}\right) + d_2 \tag{5.5}$$

However, over wide ranges of fanout, clock slew and data slew, our simulation results show that among Eq. 5.3, 5.4 and 5.5, only the model in Eq. 5.3 can fit t_{DQ}–t_{DC} over a wide range of input data slew and clock slew very well as coefficient of multiple determination can be maintained always over 0.99. To some approximation, model 5.4 or 5.5 might be acceptable. However, even Eq. 5.5 would be more accurate then traditional logarithmic function in Eq. 5.2 as shown in Fig. 5.13. The black dots are experimental results from HSPICE, the red solid line is the logarithmic fitting and the blue dash line is exponential fitting. The proposed model shows much better accuracy than the traditional.

After the latch delay model is proposed under specific fanout, clock slew and data slew, the fitting parameters in Eq. 5.3 under specific condition can be extracted and certain table can be built up. The delay in the middle

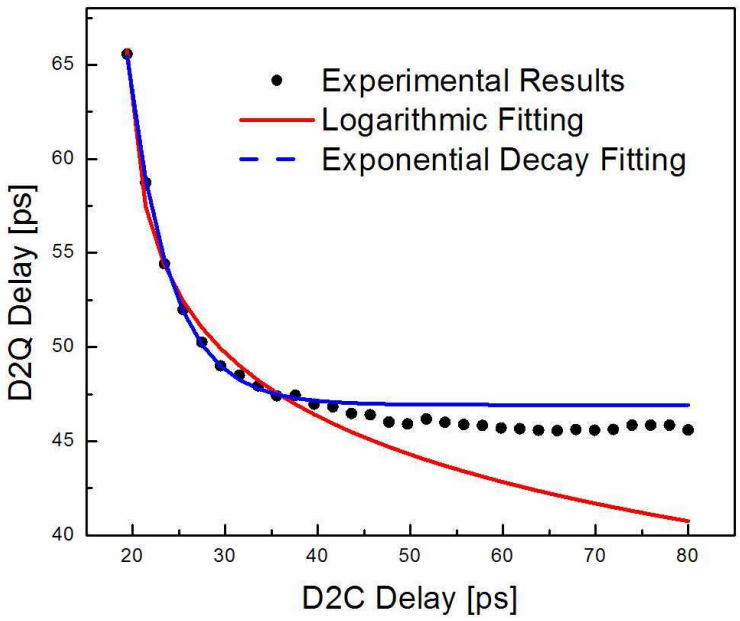


Figure 5.13: Compare exponential and logarithmic functions.
The black dots are experimental results from HSPICE, the red solid line is the logarithmic fitting and the blue dash line is exponential fitting. The proposed model shows much better accuracy than the traditional.

of nodes on the table has to be estimated. As we have several parameters such as fanout, clock slew, delay slew, the interpolation problem is formulated as follows. If f is fanout, cs as the clock slew and ds as the input data slew. We integrate these three parameters into a three dimensional vector: $\vec{w} = (f, cs, ds)$. Therefore, the multi-dimensional cubic spline interpolation is considered here. The *D2Q* delay (t_{DQ}) is a function of $\vec{w}$ and D2C delay t_{DC}, given by:

$$t_{DQ} = f(\vec{w}, t_{DC}) = a(\vec{w}) \exp\left[-b(\vec{w}) \cdot t_{DC}\right] + c(\vec{w}) \cdot t_{DC} + d(\vec{w}) \qquad (5.6)$$

where coefficients a, b, c and d are all variables of $\vec{w}$.

5.4.4 A New Latch Model for Internal Litho Variations

The statistical properties of latch is complicated, as only not the input data and clock statistical distribution but also the performance variations of the devices within the latch will affect the timing performance of a latch. According to [137], among the process, voltage and temperature variations, process variations play a significant role in the system performance. Among the process variations, CD and threshold voltage are two of the most important variation sources. As threshold voltage variations induced by dopant variations is intrinsic and effectively reduced by gate sizing [138] and systematic variations of gate length is still dominant the circuit performance [139], in this work the intra-latch variation analysis focuses on gate length variations.

As pointed out in [71], with the printability difficulty, the gate shape after lithographic process is no longer rectangular. Traditional compact device model for rectangular gate shape on layout (while the manufactured gate shape will be still non-rectangular) will introduce additional error. The key target of post-litho device compact model is to capture the difference among non-rectangular gate shapes with good compatibility to current design flow. After considering different lithographic process corner, post-litho non-rectangular device model is applied and different latch delay model parameters is extracted separately.

5.4.5 Latch Modeling in Statistical Timing Framework

In the latch based high performance circuit, timing is critical and statistical timing analysis is in strong demand. There are several works [127–129] which do statistical timing analysis for latch based circuits and the yield of critical paths and circuit can be well predicted compared with MC simulations. However, in those algorithms and MC simulations, the basic latch delay model is developed for deterministic timing analysis. In existing timing analysis, under certain loading fanout, both setup time and *D2Q* delay are fixed over different clock slew and data slew. As *D2Q* delay and setup time are constant under certain loading fanout, we have:

$$p_Q(t_Q) = \begin{cases} p_D(t_Q - t_{DQ}) & t_Q < t_C + t_{D2Q} - T_{setup} \\ 0 & t_Q > t_C + t_{D2Q} - T_{setup} \end{cases} \tag{5.7}$$

where $p_Q(t_Q)$ is the delay distribution of latch output Q, $p_D(x)$ is input data delay distribution, and t_{DQ} is *D2Q* delay. t_C is the clock delay and T_setup is setup time. From probability density function (PDF) in Eq. 5.7, cumulative distribution function (CDF) for each Q delay and final CDF can be calculated.

However, in our proposed latch delay model, there is no need to calculate specific setup time and the *D2Q* delay is just a function of *D2C* delay. Thus, the *D2Q* delay distribution will be:

$$p_{D2Q}(t_{DQ}) = p_{D2C}(g(t_{DQ})) \cdot g'(t_{DQ}) \tag{5.8}$$

where $g(x)$ is the inverse function of Eq. 5.3. If Eq. 5.5 is used for approximation, and data delay distributions are normal and clock delay is fixed at its

mean value.

$$t_Q = t_D + t_{DQ} = t_D + a_2 \cdot \exp\left(-b_2 \cdot (t_C - t_D)\right) + d_2 \tag{5.9}$$

And the final Q delay distribution should be:

$$\begin{aligned} F_Q(t_Q) = & \int_{-\infty}^{+\infty} \frac{1}{2\sqrt{2\pi}\sigma_D} \exp\left[-\frac{(t_D - \mu_D)^2}{2\sigma_D^2}\right] \\ & \cdot \left\{1 - erf\left[\frac{t_D - \mu_c - \ln\left((t_Q - t_D - d_2)/a_2\right)/b_2}{\sqrt{2}}\right]\right\} \cdot dt_D \end{aligned} \tag{5.10}$$

where

$$p_Q(t_Q) = \frac{dF_Q(t_Q)}{dt_Q}$$

$$erf(x) = 2/\sqrt{\pi} \cdot \int_0^x \exp\left(-t^2\right) dt$$

Obviously, such a distribution in Eq. 5.10 is different from the normal distribution in Eq. 5.7. The experimental results in the follow section would show the above difference.

5.5 Experimental Results of Statistical Timing with Latch

Over a very wide range (fanout: 1 16; clock slew: 5 100ps, data slew: 5 100ps), our proposed latch delay model Eq. 5.3 can fit the HSPICE simulation results with very high accuracy (coefficients of multiple determination are always over 0.99). Therefore, in the following discussions and simulations, our proposed model will be regarded as golden model. we use a typical circuit, e.g., benchmark s27[140] is used for post-latch SSTA. All other circuits have similar results.

5.5.1 The Impact of Clock Slew and Data Slew

As discussed in previous parts, not only *D2C* delay but also input data slew, clock slew and fanout can affect the *D2Q* delay. Fig. 5.14, 5.15 show the simulation results of *D2Q* minimum delay variations caused by the above external input and clock slews.

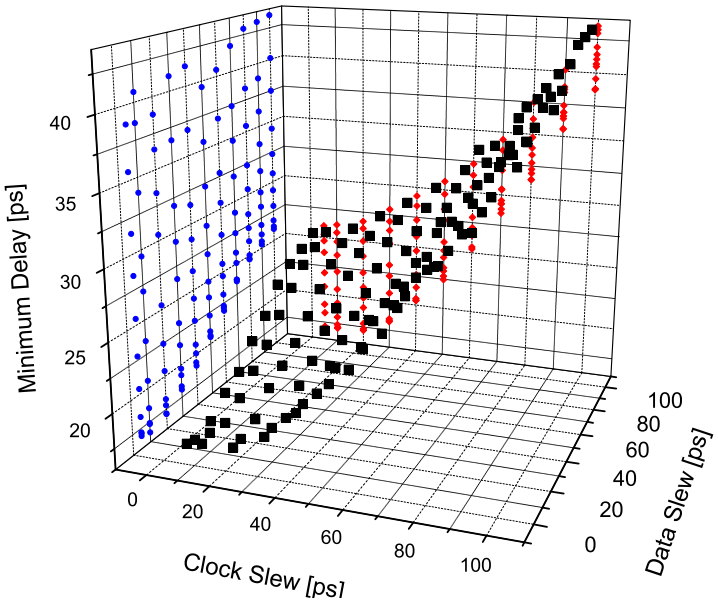


Figure 5.14: Minimum Delays' dependency on clock/input data slews.
The black square dots are latch's minimum delays at different clock slews and data slews when fanout is 4; the blue round points are projection on the plane of minimum delays and data slews; the red diamond points are projection on the plane of minimum delays and clock slews.

Fig. 5.14 shows that the minimum *D2Q* delays (among different *D2C* delays) depend on clock slews and data slews. The latch fanout of the latch is fixed to 4. The black square dots in Fig. 5.14 are minimum delays of the latch at different clock slews and data slews; the blue round points are projection on

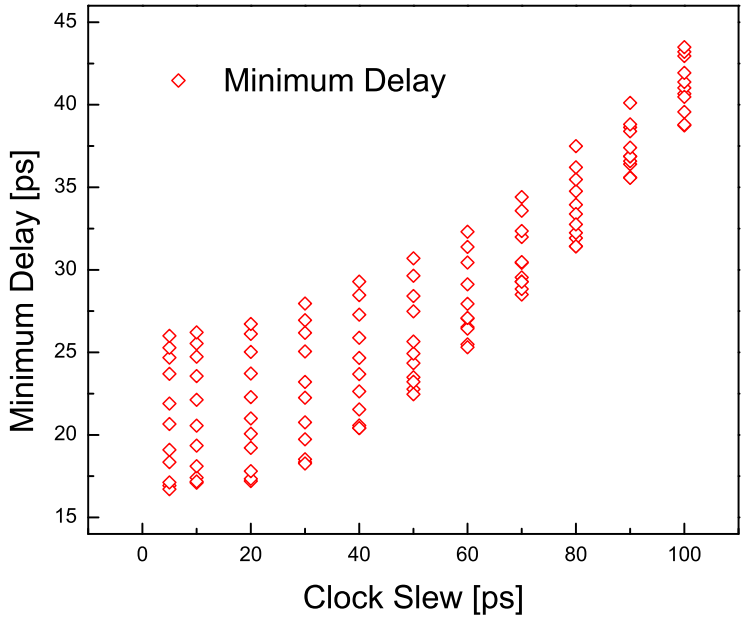


Figure 5.15: Minimum Delays' dependency on clock slews.

plane of minimum delays and data slews; the red diamond points are projection on the plane of minimum delays and clock slews. From the figures, under different clock slews and data slews, the *D2Q* delays vary over 20ps. As the overall minimum *D2Q* delay is less than 20ps, such variation range is about 100%.

Red diamond points in Fig. 5.15 are projection of black square points in Fig. 5.14 on the plane of minimum delays and clock slews. From Fig. 5.15, even under the same clock slew, the input data slew can cause about 10ps *D2Q* delay variations.

Fig. 5.16 shows the setup time dependency on data and clock slews. The variations of setup times can be over 15ps and show strong dependency

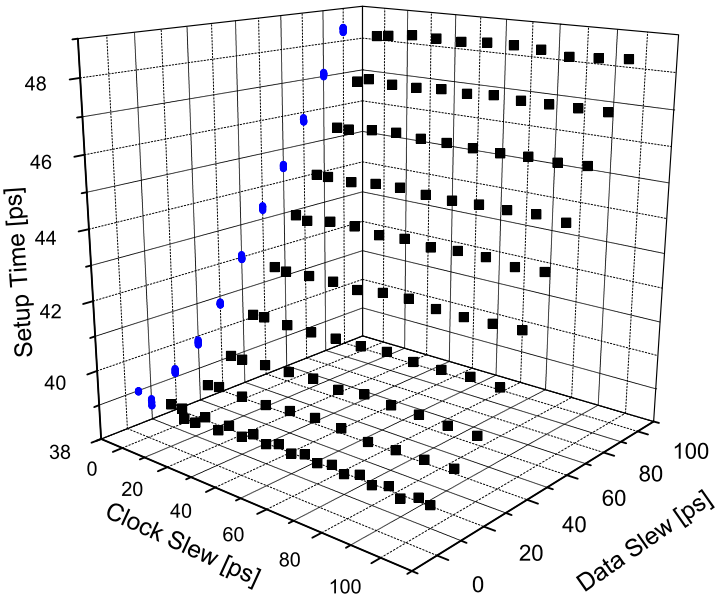


Figure 5.16: Setup times' dependency on clock slews and input data slews.

on input data slews. With limited range of data slews, the relationship is close to linearity as shown in Fig. 5.17.

More simulations show that external variations, such as data slew, clock slew, fanout, have big effect on *D2Q* delay (as shown in Fig. 5.18). During SSTA, these factors can not be omitted.

5.5.2 Statistical Timing Settings Based on MC Simulation

For benchmark s27[140], after gate sizing, Monte Carlo (MC) simulation of gate length and threshold variations are done on the most critical path made up of "NAND2 ⇒ INV1 ⇒ NOR2 ⇒ INV ⇒ NAND2 ⇒ NOR2 ⇒ NOR2". The delay and slope results are shown in Fig. 5.19.

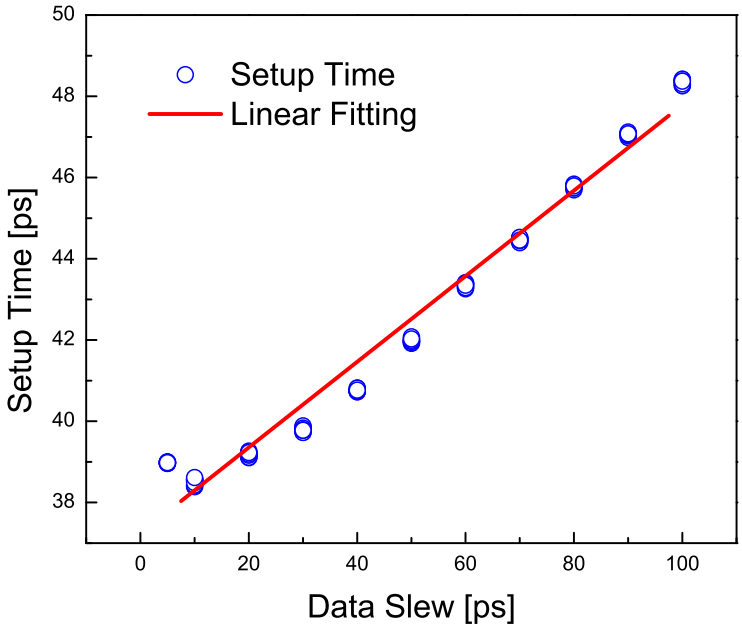


Figure 5.17: Setup times' dependency on input data slews.
Observe that the setup times are strongly dependent on data slews, and the relationship is close to linearity.

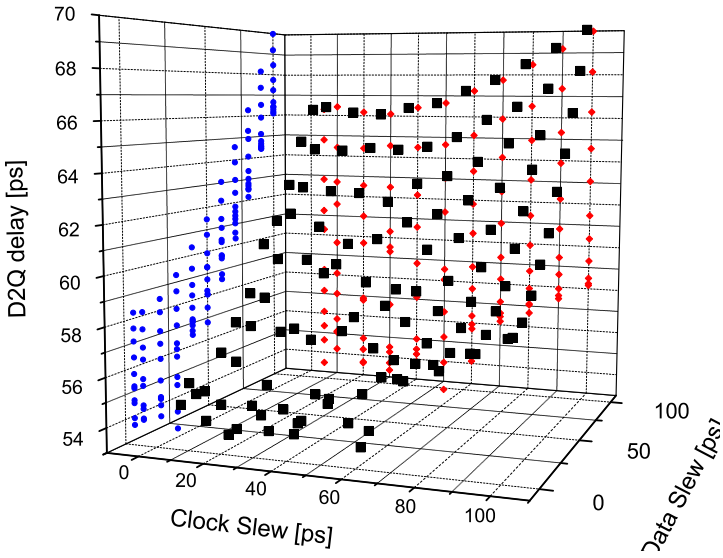


Figure 5.18: *D2Q* delay's dependency on clock slews and input data slews.

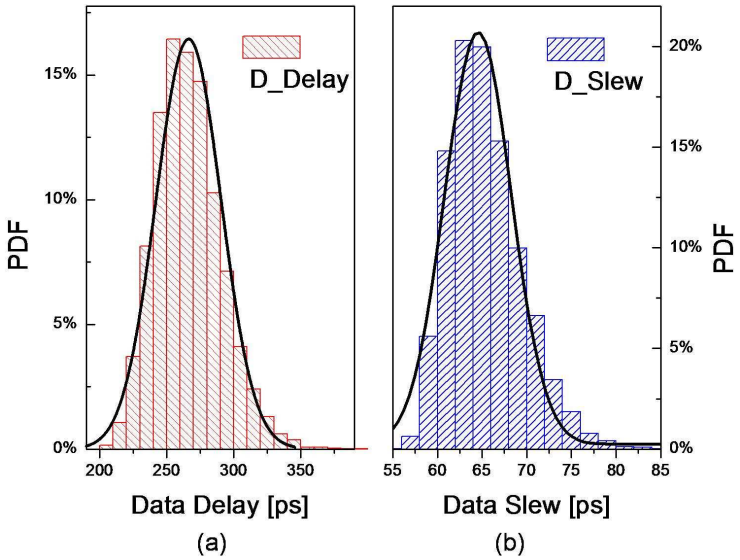


Figure 5.19: The delay and slew distribution of input data.
This is the MC simulation result of the most critical path in benchmark s27.

The mean of delay is 266.3ps with standard deviation of 24.3ps (9.1% of mean). The mean of slew is 65.4ps while the standard deviation is 4.1ps (6.3% of mean). The above results were obtained from 10,000 MC simulations. The relative standard deviation of slew is much smaller than that of delay. One intuitive explanation is that a path delay is a simple addition to gate delays while the output slew gets regenerated at each gate in the path. Thus the slew gets corrected at the output of each gate and the variation is reduced as the logic depth increases. An implied result is that the delay and slew might not be highly correlated which was verified from our MC simulations. We found that the correlation between delay and slew was 0.79 for the path mentioned in the beginning of this part.

In Fig. 5.19, the black lines represent the normal distribution fitting of delay and slew. Compared with slew, the delay distribution is more close to normal distribution. However, for approximation, it may be acceptable to use normal distribution for timing analysis.

In this part, the MC simulation results are directly sent to latch as external variations on data input terminal. The variations of clock delay and slew are omitted.

Fig. 5.20, 5.21, 5.22 and 5.23 show the simulation results and compares the Q delay distribution difference between the traditional way Eq. 5.7 (heaviside step function) used in [127–129] and our proposed model. The probability here is the output Q passed latch relative to the input signals.

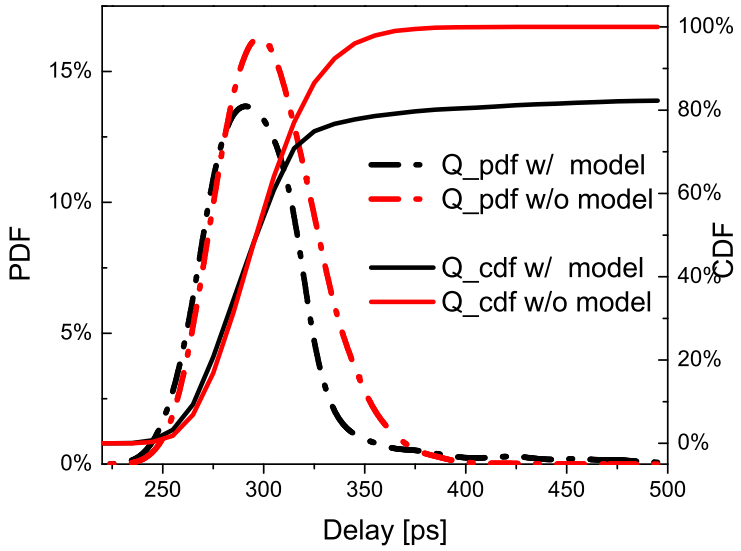


Figure 5.20: Q delay distribution based on MC simulation results. Clock period is 300ps and fanout is 2.

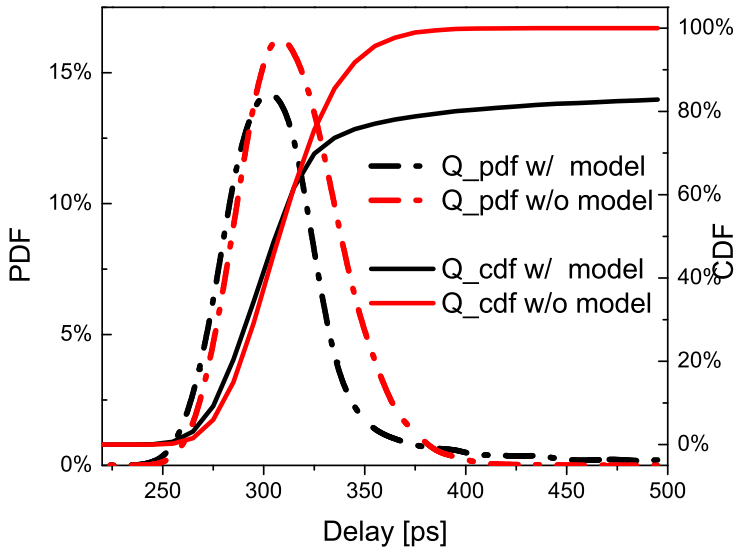


Figure 5.21: Q delay distribution based on MC simulation results. Clock period is 300ps and fanout is 4.

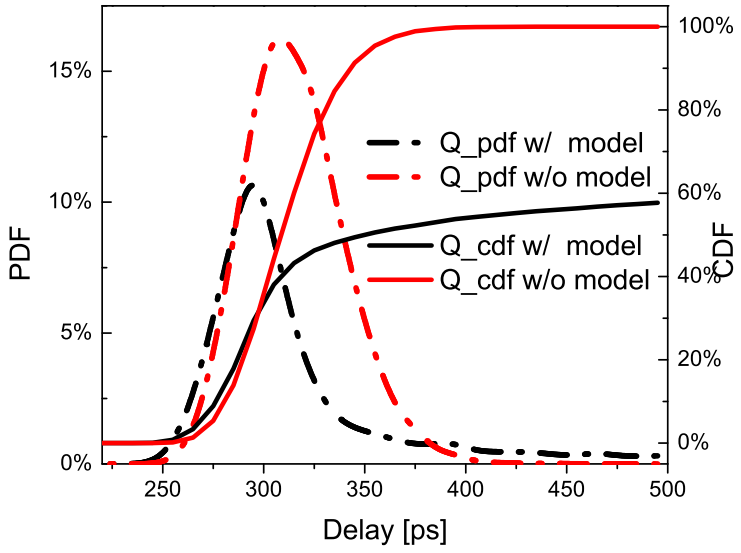


Figure 5.22: Q delay distribution based on MC simulation results. Clock period is 280ps and fanout is 4.

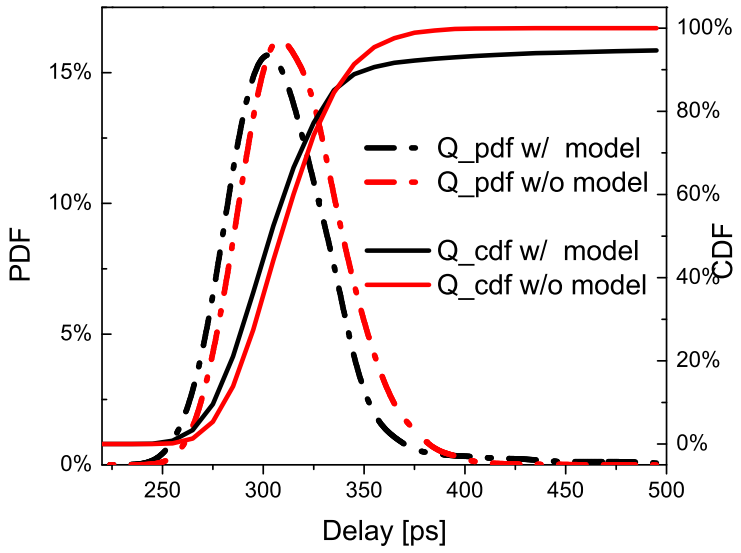


Figure 5.23: Q delay distribution based on MC simulation results. Clock period is 320ps and fanout is 4.

In Fig. 5.20, 5.21, 5.22 and 5.23, Q delay distributions based on MC simulation results are presented. The red lines are traditional output delay distribution of latch while the black lines are calculated according to our accurate latch model. The slew variations of clock and input data are omitted. They are only different in clock period (*D2C* delay) and fanout.

As the red lines are calculated from traditional latch model and step function Eq. 5.7, they are marked as "w/o model" rather than the proposed latch delay model. As practically setting of setup time is tricky, all input signals will be able to pass the latch and a constant minimum latch delay is added. After applying setup time, the output PDF and CDF would be cut when *D2C* is smaller than the setup time, yet the first part before cutting of PDF and CDF curves would be the same.

For the black line in the figures (proposed model and functional *D2Q* delays in Eq. 5.3), the probability is also the output Q passed latch relative to the input signals. Some input signals will fail to pass the latch as input data is too close to clock signals. Consequently in the figures, the black solid lines for CDF of the proposed model would not reach 100% within the *D2Q* range smaller than 500ps. Another way to explain this is that when the input data fail to pass the latch, the *D2Q* delay would be infinite and CDF could only reach 1 at ∞ point. Only when *D2Q* delay smaller than 500ps is drawn, CDF of proposed model could not reach 1 in the figures.

In this part, the data delay and slew data are directly from MC simulations and no normal distribution approximation is adopted.

The results in Fig. 5.20 are from the fact that the latch's loading fanout is 2, and the setup time of this latch is 33.4ps and minimum D2Q delay is 33.6ps. Clock delay is 300ps and clock slew is 30ps.

In Fig. 5.21, 5.22 and 5.23 the fanout is 4, the setup time of this latch is 26.5ps, minimum *D2Q* delay is 39.9ps, clock slew is fixed at 60ps, and the clock delays are 300ps, 280ps, and 320ps respectively.

From Fig. 5.20, 5.21, 5.22 and 5.23, we can see that the PDF and CDF of output Q delay distributions are different. For example, in Fig. 5.20 the two PDFs have 20% difference at the peak. In early range, the CDF calculated based on method in previous SSTA papers is close to CDF based on our proposed accurate model. However, even within this range, the PDFs

of two methods are still different from each other. Such PDF difference will be accumulated to the delay variations of later stages during statistical timing analysis.

Fig. 5.22, 5.21 and 5.23 set the clock period to be 280ps, 300ps, and 320ps, respectively. From another point of view, this means that the slacks are increasing respectively, and the paths become less timing critical. The traditional way in previous SSTA approaches has high accuracy for statistical timing analysis for the non/less timing critical path, while for the most critical paths, the traditional model becomes less accurate and our proposed latch delay model appears to be necessary.

5.5.3 More Results Based on Normal Distribution Approximation

As shown in Fig. 5.19, the data delay and slew distributions are close to normal distributions, thus normal distribution approximation is used to see the effect of correlation between delay and slew on latch delay. The original means and standard deviations of delay and slew are reserved. Meanwhile, the mean of clock delay is 300ps with 30ps standard deviation. The mean of clock slew is 60ps with 8ps standard deviation. The simulation results are shown in Fig. 5.24, 5.25, 5.26 and 5.27.

Q delay distributions are calculated based on the normal distribution approximation. The red lines are traditional output delay distribution of latch while the black lines are calculated according to our accurate latch model. The variations of clock delay and slew are considered. Fig. 5.24, 5.25 and 5.26

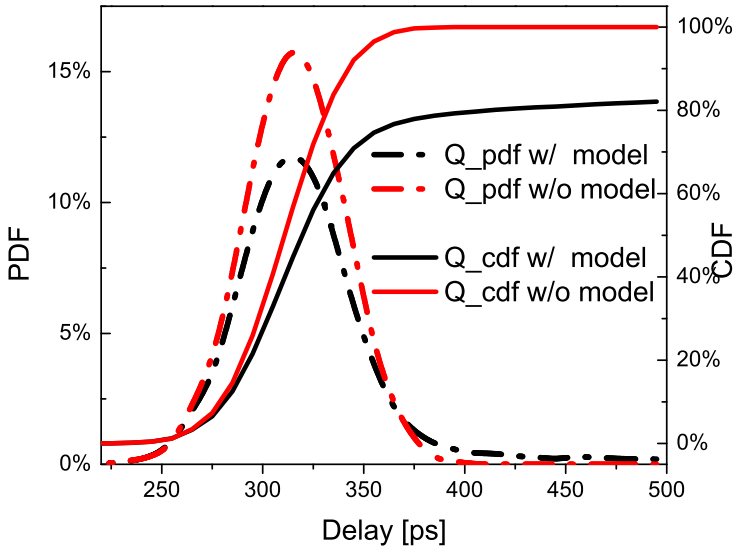


Figure 5.24: Q delay distributions.
Data delays and slews are set independently and no clock variations.

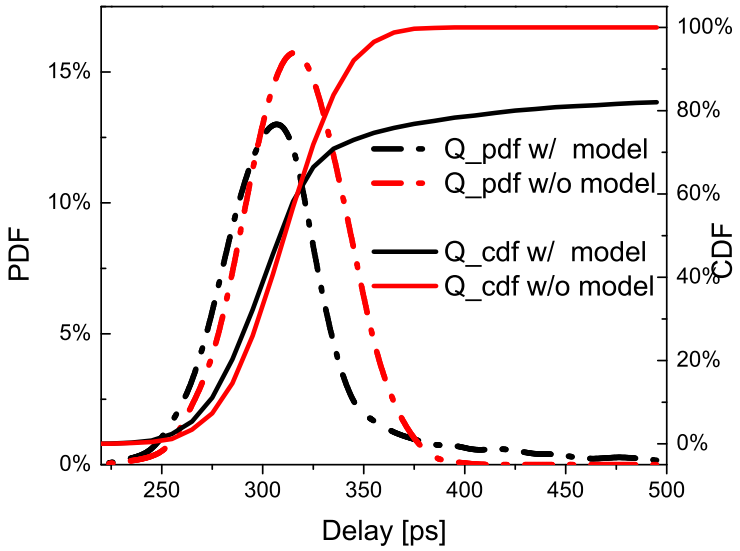


Figure 5.25: Q delay distributions.
There are no clock variations and the correlation between data delay and slew is 0.79 same to MC simulation results.

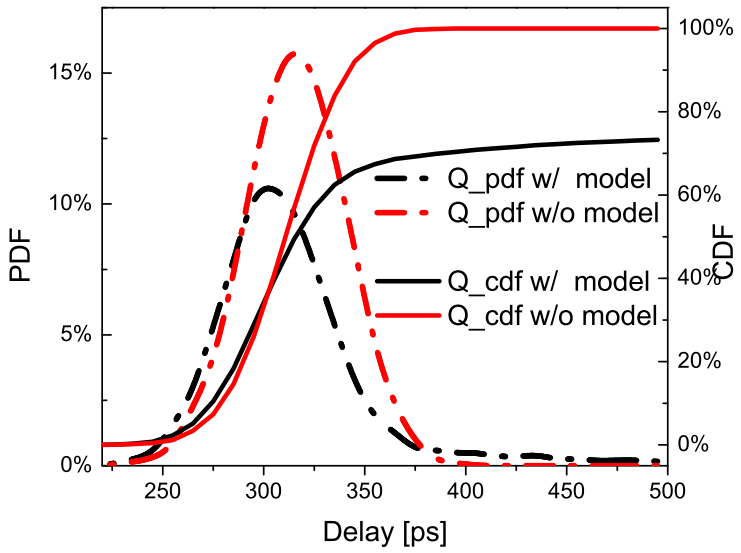


Figure 5.26: Q delay distributions.
Consider clock variations and there is 0.79 correlation between delays and slews.

are different in clock frequency and fanout. Fig. 5.27 compares PDFs of latch output based on the models of different accuracy levels.

Note that here PDF and CDF are also based on the probability of signals passing through the latch.

In Fig. 5.24, data delays and slews are set independently and no clock variations are considered. In Fig. 5.25, there is no clock variation and the correlation between data delay and slew is 0.79 as same as MC simulation results. In Fig. 5.26, the clock variations are involved with 0.79 correlation between delays and slews. Finally in Fig. 5.27, the method in previous latch SSTA papers (black line) and the condition in Fig. 5.24, Fig. 5.25 and Fig. 5.26 (the purple, red and blue line, respectively) based on the proposed model are

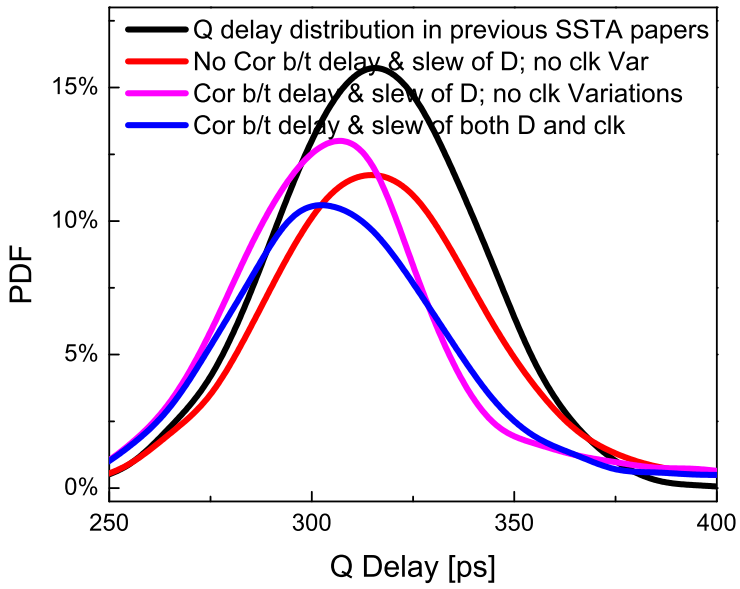


Figure 5.27: Q delay distributions.
Compares PDFs of latch output based on models of different accuracy levels. 50% error at peak is observed between rough SSTA approach compared with the proposed accurate latch delay model

compared in the PDF curves. Obvious difference can be observed. As the left side and the peak of purple line is larger than that of red line, the correlation between data delays and slews is helpful to reduce latch delays. However, with clock variations consideration, the latch delay becomes worse and about 50% error at peak is observed in previous rough SSTA approach compared with our accurate latch delay model.

5.6 Summary

We study the latch modeling for statistical timing analysis. A new perspective of latch timing is given and a much more accurate latch delay model is developed which can capture the effect of external variations of delay and slew from input data and clock. The proposed latch delay model is verified by simulations over a wide range of external variations and applied to statistical timing analysis. Compared with existing SSTA works for latch based circuits, our proposed model shows much better accuracy. The proposed comprehensive latch model is essential to accurate statistical timing analysis for combining combinational and sequential circuits.

Chapter 6

Conclusion

This dissertation studies the emerging technologies and new phenomena of nanometer scale IC. Starting from the designed layout, by modeling, optimization and simulation, the work moves ahead to mask and silicon image, reevaluates devices as well as circuits and connect layout better with silicon data to reach design and manufacturing closure.

In the manufacturing process stage (Chapter 2), a novel inverse lithography technology is proposed for sub-wavelength lithography resolution enhancement. It is computationally efficient and practically compatible to industry standard as well. In addition to developing ILT for nominal process conditions, a new dose variation aware ILT is presented. The work is verified by the newest industry $32nm$ technology and extend the state-of-art $32nm$ lithography system to $22nm$ layout. Further study should improve the local optimization of mask post processing for mask manufacturability. After developing efficient defocus models, both dose and defocus variation aware ILT would be possible if the computation complexity can be under controlled.

If the work in the lithography process stage can be seen as mask optimization and post-layout modeling, the non-rectangular device modeling card

extends the post-layout to post-litho (Chapter 3). Based on the lithography simulation results, the non-rectangular gate shapes are extracted and their effect is investigated by the proposed non-rectangular device modeling card and post-litho circuit simulation flow. This work is not only the first non-rectangular device modeling card but also has the advantage of compatibility with industry standard device models and the parameter extraction flow. Further work should also feed the more accurate simulation result back to ILT or other RET in order to produce silicon image meeting the electrical criteria. One straightforward solution is to assign the pattern edge errors with different weights based on the electrical criteria (such as timing and leakage).

Interconnect is playing a more critical role in the nanometer scale IC design especially on delay. The scattering effect of nanoscale wires is modeled for the interconnecting plan and different methods of wire sizing/shaping are discussed in Chapter 4. Based on close form resistivity model for nanometer scale Cu interconnect, new interconnect delay model and wire sizing/shaping strategies were first developed. One interesting work in future is to develop layout and mask co-optimization for metal and via layers. The optimization target should meet the timing and power requirement of interconnect and maintain both connectivity and manufacturability.

Based on the advanced modeling of process, device and interconnect, circuit level investigation is focused on statistical timing analysis with a new latch delay model in Chapter 5. For the first time, process variations effects on timing of both combinational logic and clock distribution circuits can be

integrated together through the statistical timing of latch outputs. Future direction of this work would focus on either high levels timing analysis in large system or improve the model with more mathematic work.

Appendix

Appendix 1

Derivation of Chapter 2

1.1 Derivation of Steepest Descent

The details of mathematic deduction are shown as follows,

$$
\begin{aligned}
&\frac{\partial f}{\partial \theta_p} \\
=&\frac{dm_p}{d\theta_p} \cdot \frac{\partial f}{\partial m_p} \\
=&2 \cdot \frac{dm_p}{d\theta_p} \cdot \sum_{j=1}^{MN} \left(z_j - \hat{z}_j\right) \frac{\partial}{\partial m_p} z_j \\
=&2 \cdot \frac{dm_p}{d\theta_p} \cdot \sum_{j=1}^{MN} \left(z_j - \hat{z}_j\right) \frac{\partial}{\partial m_p} S_1\left(I_j\right) \\
=&2S_2'\left(\theta_p\right) \cdot \sum_{j=1}^{MN} S_1'\left(I_j\right)\left(z_j - \hat{z}_j\right) \frac{\partial}{\partial m_p} I_j \\
=&2S_2' \cdot \sum_{j=1}^{MN} S_1'\left(z_j - \hat{z}_j\right) \frac{\partial}{\partial m_p} \left(\sum_{k=1}^{n} \sigma_k \left\| \sum_{i=1}^{MN} h_{ijk} m_i \right\|^2 \right) \\
=&\sum_{k=1}^{n} \sigma_k \cdot 2S_2' \cdot \sum_{j=1}^{MN} S_1'\left(z_j - \hat{z}_j\right) \frac{\partial}{\partial m_p} \left(\left\| \sum_{i=1}^{MN} h_{ijk} m_i \right\|^2 \right) \\
=&\sum_{k=1}^{n} \sigma_k \cdot 2S_2' \cdot \sum_{j=1}^{MN} S_1'\left(z_j - \hat{z}_j\right) \frac{\partial}{\partial m_p} \left[\left(\sum_{i=1}^{MN} h_{ijk} m_i \right) \cdot \left(\sum_{i=1}^{MN} h_{ijk} m_i \right)^* \right] \\
=&\sum_{k=1}^{n} \sigma_k \cdot 2S_2' \cdot \sum_{j=1}^{MN} S_1'\left(z_j - \hat{z}_j\right) \frac{\partial}{\partial m_p} \left[\left(\sum_{i=1}^{MN} h_{ijk} m_i \right) \cdot \left(\sum_{i=1}^{MN} h_{ijk}^* m_i \right) \right] \\
=&\sum_{k=1}^{n} \sigma_k \cdot 2S_2' \cdot \sum_{j=1}^{MN} S_1'\left(z_j - \hat{z}_j\right) \left[\left(\sum_{i=1}^{MN} h_{ijk}^* m_i \right) h_{pjk} + \left(\sum_{i=1}^{MN} h_{ijk} m_i \right) h_{pjk}^* \right] \\
=&\sum_{k=1}^{n} \sigma_k \cdot 4S_2' \cdot \sum_{j=1}^{MN} S_1'\left(z_j - \hat{z}_j\right) Real\left\{ \left(\sum_{i=1}^{MN} h_{ijk}^* m_i \right) \cdot h_{pjk} \right\} \\
=&\sum_{k=1}^{n} \sigma_k \cdot 4S_2' \cdot Real\left\{ \sum_{j=1}^{MN} \left[S_1'\left(z_j - \hat{z}_j\right) \left(\sum_{i=1}^{MN} h_{ijk} m_i \right)^* \right] \cdot h_{pjk} \right\} \\
=&\sum_{k=1}^{n} \sigma_k \cdot 4S_2' \cdot Real\left\{ \sum_{j=1}^{MN} \left[S_1'\left(z_j - \hat{z}_j\right) E_{jk}^* \right] \cdot h_{pjk} \right\}
\end{aligned}
$$

1.2 Derivation of CTR Variations

The details of mathematic deduction are shown in follows,

$$\begin{aligned}\frac{\partial}{\partial \delta_d} S_1\left(I;\alpha_1,t_r\left(\delta_d\right)\right) =& \frac{\partial}{\partial t_r} S_1\left(I;\alpha_1,t_r\right)\cdot\frac{dt_r}{d\delta_d}\\ =& -\frac{\partial}{\partial I} S_1\left(I;\alpha_1,t_r\right)\cdot\frac{dt_r}{d\delta_d}\\ =& -S_1'\cdot\frac{dt_r}{d\delta_d}\end{aligned}$$

$$\begin{aligned}f =& \sum_{j=1}^{MN}\left(z_j\left(\mathbf{M};\delta_d,0\right)-\hat{z}_j\right)^2\\ \simeq& \sum_{j=1}^{MN}\left(z_j\left(\mathbf{M};0,0\right)-\hat{z}_j\right)^2+2\delta_d\cdot\sum_{j=1}^{MN}\left(z_j\left(\mathbf{M};0,0\right)-\hat{z}_j\right)\cdot\frac{\partial}{\partial\delta_d}z_j\left(\mathbf{M};\delta_d,0\right)\\ =& f_0+2\delta_d\cdot\sum_{j=1}^{MN}\left(z_{j,0}-\hat{z}_j\right)\cdot\frac{\partial}{\partial\delta_d}S_1\left(I_j\left(\mathbf{M}\right);\alpha_1,t_r\left(\delta_d\right)\right)\\ =& f_0-2\delta_d\cdot\sum_{j=1}^{MN}S_1'\left(I_j;\alpha_1,t_{r0}\right)\cdot\left(z_{j,0}-\hat{z}_j\right)\cdot\frac{\partial}{\partial\delta_d}t_r\left(\delta_d\right)\\ =& f_0+2\delta_d\cdot\sum_{j=1}^{MN}S_1'\cdot\left(z_{j,0}-\hat{z}_j\right)\cdot t_{r0}\cdot\left(1+\delta_d\right)^{-2}\\ =& f_0+2\frac{\delta_d}{\left(1+\delta_d\right)^2}\cdot t_{r0}\cdot\sum_{j=1}^{MN}S_1'\cdot\left(z_{j,0}-\hat{z}_j\right)\end{aligned}$$

$$
\begin{aligned}
F_{\max}\left(M;\delta_{\mathbf{d}},0\right) =& f_0+2\max\left\{\left.\frac{\delta_d}{\left(1+\delta_d\right)^2}\cdot t_{r0}\cdot\sum_{j=1}^{MN}S_1'\cdot\left(z_{j,0}-\hat{z}_j\right)\right|\delta_d=\delta_{d,i}\right\}\\
=& f_0+2t_{r0}\cdot\left|\sum_{j=1}^{MN}S_1'\cdot\left(z_{j,0}-\hat{z}_j\right)\right|\cdot\max\left\{\left.\frac{\delta_d}{\left(1+\delta_d\right)^2}\right|\delta_d=\delta_{d,i}\right\}\\
\approx& f_0+\frac{2t_{r0}\cdot\delta_{d,\max}}{\left(1-\delta_{d,\max}\right)^2}\cdot\left|\sum_{j=1}^{MN}S_1'\cdot\left(z_{j,0}-\hat{z}_j\right)\right|
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial}{\partial\theta_p}F_{\max}\left(\Theta;\delta_{\mathbf{d}},0\right)\\
=&\frac{\partial f_0}{\partial\theta_p}+\frac{2t_{r0}\cdot\delta_{d,\max}}{\left(1-\delta_{d,\max}\right)^2}\cdot\frac{\partial}{\partial\theta_p}\left|\sum_{j=1}^{MN}S_1'\cdot\left(z_{j,0}-\hat{z}_j\right)\right|\\
\approx&\frac{\partial f_0}{\partial\theta_p}+\frac{2t_{r0}\cdot\delta_{d,\max}}{\left(1-\delta_{d,\max}\right)^2}\cdot\frac{\partial}{\partial\theta_p}S_3\left(\sum_{j=1}^{MN}S_1'\cdot\left(z_{j,0}-\hat{z}_j\right);\alpha_3\right)\\
=&\frac{\partial f_0}{\partial\theta_p}+\frac{2t_{r0}\cdot\delta_{d,\max}}{\left(1-\delta_{d,\max}\right)^2}\cdot\frac{dm_p}{d\theta_p}\frac{\partial}{\partial m_p}S_3\left(\sum_{j=1}^{MN}S_1'\cdot\left(z_{j,0}-\hat{z}_j\right);\alpha_3\right)\\
=&\frac{\partial f_0}{\partial\theta_p}+\frac{2t_{r0}\cdot\delta_{d,\max}}{\left(1-\delta_{d,\max}\right)^2}\cdot S_2'\cdot S_3'\cdot\frac{\partial}{\partial m_p}\left(\sum_{j=1}^{MN}S_1'\cdot\left(z_{j,0}-\hat{z}_j\right)\right)\\
=&\frac{\partial f_0}{\partial\theta_p}+\frac{2t_{r0}\cdot\delta_{d,\max}}{\left(1-\delta_{d,\max}\right)^2}\cdot S_2'\cdot S_3'\cdot\sum_{j=1}^{MN}\frac{\partial}{\partial m_p}\left(S_1'\cdot\left(z_{j,0}-\hat{z}_j\right)\right)\\
=&\frac{\partial f_0}{\partial\theta_p}+\frac{2t_{r0}\cdot\delta_{d,\max}}{\left(1-\delta_{d,\max}\right)^2}\cdot S_2'\cdot S_3'\cdot\sum_{j=1}^{MN}\frac{\partial I_{j,0}}{\partial m_p}\cdot\frac{d}{dI_{j,0}}\left(S_1'\cdot\left(z_{j,0}-\hat{z}_j\right)\right)\\
=&\frac{\partial f_0}{\partial\theta_p}+\frac{2t_{r0}\cdot\delta_{d,\max}}{\left(1-\delta_{d,\max}\right)^2}\cdot S_2'\cdot S_3'\cdot\sum_{j=1}^{MN}\left[S_1''\cdot\left(z_{j,0}-\hat{z}_j\right)+\left(S_1'\right)^2\right]\cdot\frac{\partial I_{j,0}}{\partial m_p}\\
=&2\cdot S_2'\cdot\sum_{j=1}^{MN}S_1'\cdot\left(z_{j,0}-\hat{z}_j\right)\cdot\frac{\partial I_{j,0}}{\partial m_p}\\
&+\frac{2t_{r0}\cdot\delta_{d,\max}}{\left(1-\delta_{d,\max}\right)^2}\cdot S_2'\cdot S_3'\cdot\cdot\sum_{j=1}^{MN}\left[S_1''\cdot\left(z_{j,0}-\hat{z}_j\right)+\left(S_1'\right)^2\right]\cdot\frac{\partial I_{j,0}}{\partial m_p}\\
=&2\cdot S_2'\cdot\sum_{j=1}^{MN}\left[\begin{array}{l}S_1'\cdot\left(z_{j,0}-\hat{z}_j\right)\\+\dfrac{t_{r0}\cdot\delta_{d,\max}}{\left(1-\delta_{d,\max}\right)^2}\cdot S_3'\cdot S_1''\cdot\left(z_{j,0}-\hat{z}_j\right)\\+\dfrac{t_{r0}\cdot\delta_{d,\max}}{\left(1-\delta_{d,\max}\right)^2}\cdot S_3'\cdot\left(S_1'\right)^2\end{array}\right]\cdot\frac{\partial I_{j,0}}{\partial m_p}
\end{aligned}
$$

$$
\begin{aligned}
&\nabla F_{\max}\left(M;\delta_{\mathbf{d}},0\right)\\
=&\sum_{k=1}^{n}\sigma_k\cdot 4\cdot S_2'\odot Real\left\{\left[\begin{array}{l} S_1'\odot\left(z_j-\hat{z}_j\right)\\ +\dfrac{t_{r0}\cdot\delta_{d,\max}}{\left(1-\delta_{d,\max}\right)^2}\cdot S_3'\cdot\left(\begin{array}{l} S_1''\odot\left(z_{j,0}-\hat{z}_j\right)\\ +S_1'\odot S_1'\end{array}\right)\end{array}\right]\odot E_{jk}^{*}\otimes h_{pjk}\right\}
\end{aligned}
$$

1.3 Derivation of Dose Variations

$$
\begin{aligned}
f=&\sum_{j=1}^{MN}\left(z_j\left(\mathbf{M};\delta_d,0\right)-\hat{z}_j\right)^2\\
=&\sum_{j=1}^{MN}\left(z_j\left(\mathbf{M};0,0\right)-\hat{z}_j\right)^2\\
&+2\delta_d\cdot\sum_{j=1}^{MN}\left(z_j\left(\mathbf{M};0,0\right)-\hat{z}_j\right)\cdot\frac{\partial}{\partial\delta_d}z_j\left(\mathbf{M};\delta_d,0\right)\\
=&f_0+2\delta_d\cdot\sum_{j=1}^{MN}S_1'\left(I_{j,0}\right)\left(z_j-\hat{z}_j\right)\frac{\partial}{\partial\delta_d}I_j\left(\mathbf{M};\delta_d,0\right)\\
=&f_0+2\delta_d\cdot\sum_{j=1}^{MN}S_1'\left(I_j\right)\left(z_{j,0}-\hat{z}_j\right)\frac{\partial}{\partial\delta_d}\left[I_j\left(\mathbf{M};0,0\right)\cdot\left(1+\delta_d\right)\right]\\
=&f_0+2\delta_d\cdot\sum_{j=1}^{MN}S_1'\left(I_j\right)\cdot\left(z_{j,0}-\hat{z}_j\right)\cdot I_j\left(\mathbf{M};0,0\right)\\
=&f_0+2\delta_d\cdot\sum_{j=1}^{MN}S_1'\cdot\left(z_{j,0}-\hat{z}_j\right)\cdot I_{j,0}
\end{aligned}
$$

$$
\begin{aligned}
&F_{\max}\left(M;\delta_{\mathbf{d}},0\right)\\
=&f_0+\max\left\{2\delta_d\cdot\sum_{j=1}^{MN}S_1'\cdot\left(z_{j,0}-\hat{z}_j\right)\cdot I_{j,0}\middle|\,\delta_d=\delta_{d,i}\right\}\\
=&f_0+2\cdot\left|\sum_{j=1}^{MN}S_1'\cdot\left(z_{j,0}-\hat{z}_j\right)\cdot I_{j,0}\right|\cdot\max\left\{\delta_d|\,\delta_d=\delta_{d,i}\right\}\\
=&f_0+2\delta_{d,\max}\cdot\left|\sum_{j=1}^{MN}S_1'\cdot\left(z_{j,0}-\hat{z}_j\right)\cdot I_{j,0}\right|
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial}{\partial\theta_p}F_{\max}\left(\Theta;\delta_{\mathbf{d}},0\right)\\
=&\frac{\partial f_0}{\partial\theta_p}+2\delta_{d,\max}\cdot\frac{\partial}{\partial\theta_p}\left|\sum_{j=1}^{MN}S_1'\cdot\left(z_{j,0}-\hat{z}_j\right)\cdot I_{j,0}\right|\\
\cong&\frac{\partial f_0}{\partial\theta_p}+2\delta_{d,\max}\cdot\frac{\partial}{\partial\theta_p}S_3\left(\sum_{j=1}^{MN}S_1'\cdot\left(z_{j,0}-\hat{z}_j\right)\cdot I_{j,0};\alpha_3\right)\\
=&\frac{\partial f_0}{\partial\theta_p}+2\delta_{d,\max}\cdot\frac{dm_p}{d\theta_p}\frac{\partial}{\partial m_p}S_3\left(\sum_{j=1}^{MN}S_1'\cdot\left(z_{j,0}-\hat{z}_j\right)\cdot I_{j,0};\alpha_3\right)\\
=&\frac{\partial f_0}{\partial\theta_p}+2\delta_{d,\max}\cdot S_2'\cdot S_3'\cdot\frac{\partial}{\partial m_p}\left(\sum_{j=1}^{MN}S_1'\cdot\left(z_{j,0}-\hat{z}_j\right)\cdot I_{j,0}\right)\\
=&\frac{\partial f_0}{\partial\theta_p}+2\delta_{d,\max}\cdot S_2'\cdot S_3'\cdot\sum_{j=1}^{MN}\frac{\partial}{\partial m_p}\left(S_1'\cdot\left(z_{j,0}-\hat{z}_j\right)\cdot I_{j,0}\right)\\
=&\frac{\partial f_0}{\partial\theta_p}+2\delta_{d,\max}\cdot S_2'\cdot S_3'\cdot\sum_{j=1}^{MN}\frac{\partial I_{j,0}}{\partial m_p}\cdot\frac{d}{dI_{j,0}}\left(S_1'\cdot\left(z_{j,0}-\hat{z}_j\right)\cdot I_{j,0}\right)\\
=&\frac{\partial f_0}{\partial\theta_p}+2\delta_{d,\max}\cdot S_2'\cdot S_3'\cdot\\
&\sum_{j=1}^{MN}\left[S_1''\cdot\left(z_{j,0}-\hat{z}_j\right)\cdot I_{j,0}+\left(S_1'\right)^2\cdot I_{j,0}+S_1'\cdot\left(z_{j,0}-\hat{z}_j\right)\right]\cdot\frac{\partial I_{j,0}}{\partial m_p}\\
=&2\cdot S_2'\cdot\sum_{j=1}^{MN}S_1'\cdot\left(z_{j,0}-\hat{z}_j\right)\cdot\frac{\partial I_{j,0}}{\partial m_p}+2\delta_{d,\max}\cdot S_2'\cdot S_3'\cdot\\
&\cdot\sum_{j=1}^{MN}\left[S_1''\cdot\left(z_{j,0}-\hat{z}_j\right)\cdot I_{j,0}+\left(S_1'\right)^2\cdot I_{j,0}+S_1'\cdot\left(z_{j,0}-\hat{z}_j\right)\right]\cdot\frac{\partial I_{j,0}}{\partial m_p}\\
=&2\cdot S_2'\cdot\sum_{j=1}^{MN}\begin{bmatrix}S_1'\cdot\left(z_{j,0}-\hat{z}_j\right)\\+\delta_{d,\max}\cdot S_3'\cdot S_1''\cdot\left(z_{j,0}-\hat{z}_j\right)\cdot I_{j,0}\\+\delta_{d,\max}\cdot S_3'\cdot\left(S_1'\right)^2\cdot I_{j,0}\\+\delta_{d,\max}\cdot S_3'\cdot S_1'\cdot\left(z_{j,0}-\hat{z}_j\right)\end{bmatrix}\cdot\frac{\partial I_{j,0}}{\partial m_p}
\end{aligned}
$$

$$\nabla F_{\max}\left(M;\delta_{\mathbf{d}},0\right)$$

$$=\sum_{k=1}^{n}\sigma_k\cdot 4\cdot S_2'\odot Real\left\{\left[\begin{array}{l} S_1'\odot\left(z_j-\hat{z}_j\right) \\ +\delta_{d,\max}\cdot S_3'\cdot \\ \cdot\left(\begin{array}{l} S_1''\odot\left(z_{j,0}-\hat{z}_j\right)\odot I_{j,0} \\ +S_1'\odot S_1'\odot I_{j,0} \\ +S_1'\odot\left(z_{j,0}-\hat{z}_j\right) \end{array}\right) \end{array}\right]\odot E_{jk}^{*}\otimes h_{pjk}\right\}$$

Bibliography

[1] J. Rabaey, A. Chandrakasan, and B. Nikolić, *Digital integrated circuits: a design perspective*. Prentice hall New Jersey, 2nd ed., 2000.

[2] M. Abramovici and P. Bradley, "A new approach to in-system silicon validation and debug," *EDA DesignLine*, Sept. 16, 2007.

[3] "International technology roadmap for semiconductors (ITRS)," 2003.

[4] K. A. Dunn, "Reliability implications of ultra-thin copper lines and low-k dielectrics." presented on International SEMATECH, October 2004.

[5] W. Steinhoegl, G. Schindler, G. Steinlesberger, and M. Engelhardt, "Size-dependent resistivity of metallic wires in the mesoscopic range," *Physical Review B (Condensed Matter and Materials Physics)*, vol. 66, no. 7, p. 075414, 2002.

[6] S. M. Rossnagel and T. S. Kuan, "Alteration of cu conductivity in the size effect regime," *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures*, vol. 22, no. 1, p. 240, 2004.

[7] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, 1965.

[8] T. R. Halfhill, "Intel defends x86 strategy: desktop PCs are still important, but mobile computing is crucial," *Microprocessor Report*, Aug. 24 2009.

[9] A. Chakraborty, S. X. Shi, and D. Z. Pan, "Layout level timing optimization by leveraging active area dependent mobility of strained-silicon devices," in *Proceedings of the conference on Design, automation and test in Europe*, pp. 849–855, ACM, 2008.

[10] S. Wolf and R. N. Tauber, *Silicon Processing for the VLSI Era, Vol. 1: Process Technology*. Lattice Press, 1999.

[11] C. A. Mack, *Fundamental principles of optical lithography : the science of microfabrication*. Wiley, January 2007.

[12] C. A. Mack, "IEEE Spectrum: Seeing double," *IEEE Spectrum*, pp. 47–51, November 2008.

[13] A. K.-K. Wong, *Resolution Enhancement Techniques in Optical Lithography*. Bellingham, WA: SPIE, 2001.

[14] *Taurus-OPC MODULES, Optical Lithography Correction, User's Manual*, 1999.

[15] N. Cobb, "Flexible sparse and dense OPC algorithms," in *SPIE on Photomask and Next-Generation Lithography Mask Technology XII*, vol. 5853, pp. 693–702, 2005.

[16] N. Cobb and D. Dudau, "Dense OPC and verification for 45nm," in *SPIE on Optical Microlithography XIX*, vol. 6154, pp. 191–196, 2006.

[17] T. Lin, F. Robert, A. Borjon, G. Russell, C. Martinelli, A. Moore, and Y. Rody, "SRAF placement and sizing using inverse lithography technology," vol. 6520, p. 65202A, SPIE, 2007.

[18] J.-C. Yu, P. Yu, and H.-Y. Chao, "Model-based sub-resolution assist features using an inverse lithography method," vol. 7140, p. 714014, SPIE, 2008.

[19] F. M. Schellenberg, "Resolution enhancement technology: the past, the present, and extensions for the future," vol. 5377, pp. 1–20, SPIE, 2004.

[20] S. I. Sayegh and B. E. A. Saleh, "Image design: generation of a prescribed image at the output of a band-limited system," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PAMI-5, pp. 441–445, July 1983.

[21] S. Sayegh, B. Saleh, and K. Nashold, "Image design: generation of a prescribed image through a diffraction-limited system with high-contrast recording," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, pp. 460–465, Apr 1985.

[22] Y. Liu and A. Zakhor, "Binary and phase shifting mask design for optical lithography," *Semiconductor Manufacturing, IEEE Transactions on*, vol. 5, pp. 138–152, May 1992.

[23] X. Ma and G. Arce, "Binary mask optimization for inverse lithography with partially coherent illumination," *J. Opt. Soc. Am. A*, vol. 25, no. 12, pp. 2960–2970, 2008.

[24] Y.-H. Oh, J.-C. Lee, and S. Lim, "Resolution enhancement through optical proximity correction and stepper parameter optimization for 0.12-mu m mask pattern," vol. 3679, pp. 607–613, SPIE, 1999.

[25] V. Singh, B. Hu, K. Toh, S. Bollepalli, S. Wagner, and Y. Borodovsky, "Making a trillion pixels dance," vol. 6924, p. 69240S, SPIE, 2008.

[26] A. Poonawala and P. Milanfar, "Mask design for optical microlithographyan inverse imaging problem," *Image Processing, IEEE Transactions on*, vol. 16, pp. 774–788, March 2007.

[27] A. E. Rosenbluth, S. Bukofsky, C. Fonseca, M. Hibbs, K. Lai, A. F. Molless, R. N. Singh, and A. K. K. Wong, "Optimum mask and source patterns to print a given shape," *Journal of Microlithography, Microfabrication, and Microsystems*, vol. 1, no. 1, pp. 13–30, 2002.

[28] N. Cobb and Y. Granik, "New concepts in OPC," vol. 5377, pp. 680–690, SPIE, 2004.

[29] Y. Cao, Y.-W. Lu, L. Chen, and J. Ye, "Optimized hardware and software for fast full-chip simulation," vol. 5754, pp. 407–414, SPIE, 2004.

[30] F. M. Schellenberg, O. Toublan, L. Capodieci, and B. Socha, "Adoption of OPC and the impact on design and layout," in *DAC '01: Proceed-*

ings of the 38th annual Design Automation Conference, (New York, NY, USA), pp. 89–92, ACM, 2001.

[31] H. H. Hopkins, "On the diffraction theory of optical images," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 217, no. 1130, pp. 408–432, 1953.

[32] M. Born and E. Wolf, *Principles of Optics : Electromagnetic Theory of Propagation, Interference and Diffraction of Light.* Oxford: Cambridge University Press, 7th ed., Oct. 1999.

[33] N. B. Cobb, *Fast optical and process proximity correction algorithms for integrated circuit manufacturing.* PhD thesis, University of California, Berkeley, 1998. Chair-Zakhor, Avideh.

[34] J. Randall, K. G. Ronse, T. Marschner, A.-M. Goethals, and M. Ercken, "Variable-threshold resist models for lithography simulation," vol. 3679, pp. 176–182, SPIE, 1999.

[35] W.-C. Huang, C.-H. Lin, C.-C. Kuo, C. C. Huang, J. F. Lin, J.-H. Chen, R.-G. Liu, Y. C. Ku, and B.-J. Lin, "Two threshold resist models for optical proximity correction," vol. 5377, pp. 1536–1543, SPIE, 2004.

[36] D. Fuard, M. Besacier, and P. Schiavone, "Validity of the diffused aerial image model: an assessment based on multiple test cases," in *Proceedings of the SPIE* (A. Yen, ed.), vol. 5040, pp. 1536–1543, SPIE, 2003.

[37] Mentor Graphics Corporation, *Calibre Manual*, software version 2007.2.15 ed., Feb. 2007.

[38] F. Liu and S. X. Shi, "CORE: computational optimization for resolution enhancement in sub-wavelength lithography," *submitted to DAC*, 2010.

[39] S. Sherif, B. Saleh, and R. De Leone, "Binary image synthesis using mixed linear integer programming," *Image Processing, IEEE Transactions on*, vol. 4, pp. 1252–1257, Sep 1995.

[40] M. Frigo and S. G. Johnson, "The design and implementation of FFTW3," *Proceedings of the IEEE*, vol. 93, no. 2, pp. 216–231, 2005. Special issue on "Program Generation, Optimization, and Platform Adaptation".

[41] R. A. Ferguson, M. A. Lavin, L. W. Liebmann, and A. K. Wong, "Process window based optical proximity correction of lithographic images." US Patent 6578190, Jan 2001. US Patent 6578190.

[42] P. Yu, S. X. Shi, and D. Z. Pan, "True process variation aware optical proximity correction with variational lithography modeling and model calibration," *The Journal of Microlithography, Microfabrication, and Microsystems (JM3)*, vol. Special Edition of Resolution Enhancement Techniques and Design for Manufacturability, September 2007.

[43] R. Lugg, M. StJohn, Y. Zhang, A. Yang, and P. Van Adrichem, "Full-chip process window aware OPC capability assessment," in *Photomask*

Technology 2007. Edited by Naber, Robert J.; Kawahira, Hiroichi. Proceedings of the SPIE, Volume 6730, pp. 67302U (2007)., vol. 6730 of *Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*, Oct. 2007.

[44] Q. Zhang, Q. Yan, Y. Zhang, and K. Lucas, "Continuous process window modeling for process variation aware OPC and lithography verification," in *Design for Manufacturability through Design-Process Integration II. Edited by Singh, Vivek K.; Rieger, Michael L. Proceedings of the SPIE, Volume 6925, pp. 69251A-69251A-10 (2008).*, vol. 6925 of *Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*, Apr. 2008.

[45] M. Dusa, R. Moerman, B. Singh, P. Friedberg, R. Hoobler, and T. Zavecs, "Intra-wafer cdu characterization to determine process and focus contributions based on scatterometry metrology," in *Proceedings of the SPIE* (K. W. Tobin and Jr., eds.), vol. 5378, pp. 93–104, SPIE, 2004.

[46] M. Mani, A. K. Sing, and M. Orshansky, "Joint design-time and post-silicon minimization of parametric yield loss using adjustable robust optimization," in *ICCAD '06: Proceedings of the 2006 IEEE/ACM international conference on Computer-aided design*, (New York, NY, USA), pp. 19–26, ACM, 2006.

[47] M. Mani, A. Devgan, and M. Orshansky, "An efficient algorithm for statistical minimization of total power under timing yield constraints," in

DAC '05: Proceedings of the 42nd annual Design Automation Conference, (New York, NY, USA), pp. 309–314, ACM, 2005.

[48] K. Madsen, "An algorithm for minimax solution of overdetermined systems of non-linear equations," *IMA J Appl Math*, vol. 16, no. 3, pp. 321–328, 1975.

[49] J. Hald and K. Madsen, "Combined LP and quasi-newton methods for minimax optimization," *Journal of Mathematical Programming*, vol. 20, pp. 49–62, Dec. 1981.

[50] J. W. Bandler, W. Kellermann, and K. Madsen, "A superlinearly convergent minimax algorithm for microwave circuit design," *IEEE Transactions on Microwave Theory Techniques*, vol. 33, pp. 1519–1530, Dec. 1985.

[51] M. LaPedus, "IBM sees immersion at 22nm, pushes out EUV," *EE-Times*, Feb. 23, 2007.

[52] A. K.-K. Wong, "Microlithography: trends, challenges, solutions, and their impact on design," *IEEE Micro*, vol. 23, no. 2, pp. 12–21, 2003.

[53] L. W. Liebmann, "Layout impact of resolution enhancement techniques: impediment or opportunity?," in *International Symposium on Physical Design*, (Monterey, CA, USA), pp. 110–117, 2003.

[54] J. Yang, L. Capodieci, and D. M. Sylvester, "Advanced timing analysis based on post-OPC extraction of critical dimensions," in *Design*

Automation Conference, p. 359364, ACM Press(New York, NY, USA), 2005.

[55] J. Chung, M. Jeng, J. E. Moon, A. T. Wu, T. Y. Chan, P. K. Ko, and C. Hu, "Deep-submicrometer MOS device fabrication using a photoresist-ashing technique," *Electron Device Letters*, vol. 9, no. 4, pp. 186–188, 1988.

[56] K. Asano, Y.-K. Choi, T.-J. King, and C. Hu, "Patterning sub-30-nm MOSFET gate with i-line lithography," *Electron Devices, IEEE Transactions on*, vol. 48, pp. 1004–1006, May 2001.

[57] C.-Y. Sin, B.-H. Chen, W. L. Loh, J. Yu, P. Yelehanka, A. See, and L. Chan, "Resist trimming in high-density cf_4/o_2 plasmas for sub-0.1 μm device fabrication," *Journal of vacuum science and technology B, Microelectronics and nanometer structures*, vol. 20, no. 5, pp. 1974–1981, 2002.

[58] R. Brumester, J. Winnerl, and F. Neppl, "Realization of deep-submicron MOSFETS by lateral etching," *Microelectronic Engineering*, vol. 13, no. 1-4, pp. 473–476, 1991.

[59] W. L. Wei, H. Qizhi, M. Hanratty, D. Rogers, A. Chatterjee, R. Kraft, and R. A. Chapman, "Fabrication of 0.06 μm poly-si gate using DUV lithography with a designed $Si_xO_yN_z$ film as an ARC and hardmask," in *Symposium on VLSI Technology*, pp. 131–132, 1997.

[60] H. Fukutome, T. Aoyama, Y. Momiyama, T. Kubo, Y. Tagawa, and H. Arimoto, "Direct evaluation of gate line edge roughness impact on extension profiles in sub-50nm N-MOSFETs," in *IEEE International Electron Devices Meeting*, p. 433, 2004.

[61] J. A. Croon, G. Storms, S. Winkelmeier, I. Pollentier, M. Ercken, S. Decoutere, W. Sansen, and H. E. Maes, "Line edge roughness: characterization, modeling and impact on device behavior," in *IEEE International Electron Devices Meeting*, pp. 307–310, 2002.

[62] T. Linton, M. Chandhok, B. J. Rice, and G. Schrom, "Determination of the line edge roughness specification for 34nm devices," in *IEEE International Electron Devices Meeting*, pp. 303–306, 2002.

[63] K. Seong-Dong, H. Wada, and J. C. S. Woo, "TCAD-based statistical analysis and modeling of gate line-edge roughness effect on nanoscale mos transistor performance and scaling," *Transactions Semiconductor Manufacturing*, vol. 17, no. 2, p. 192, 2004. 0894-6507.

[64] S. Xiong and J. Bokor, "A simulation study of gate line edge roughness effects on doping profiles of short-channel MOSFET devices," *IEEE Transactions on Electron Devices*, vol. 51, no. 2, p. 228, 2004. 0018-9383.

[65] K. Cao, S. Dobre, and J. Hu, "Standard cell characterization considering lithography induced variations," in *Design Automation Conference*, pp. 801–804, 2006.

[66] F.-L. Heng, J.-F. Lee, P. Gupta, and L. W. Liebmann, "Toward through-process layout quality metrics," in *SPIE on Design and Process Integration for Microelectronic Manufacturing III*, vol. 5756, pp. 161–167, 2005.

[67] W. J. Poppe, L. Capodieci, J. Wu, and A. Neureuther, "From poly line to transistor: building BSIM models for non-rectangular transistors," in *SPIE on Design and Process Integration for Microelectronic Manufacturing IV*, vol. 6156, pp. 235–243, 2006.

[68] "BSIM website," *http://www-device.eecs.berkeley.edu/ bsim3/*.

[69] "PTM website," *http://www.eas.asu.edu/ ptm/*.

[70] C.-T. Sah, "A history of MOS transistor compact modeling," in *Workshop Compact Modeling*, (Anaheim, CA: Nanotech Science and Technology Institute), pp. 347–390, 2005.

[71] S. X. Shi, P. Yu, and D. Z. Pan, "A unified non-rectangular device and circuit simulation model for timing and power," in *ICCAD '06: Proceedings of the 2006 IEEE/ACM international conference on Computer-aided design*, (New York, NY, USA), pp. 423–428, ACM, 2006.

[72] "PSP website," *http://pspmodel.asu.edu/*.

[73] N. Arora, *MOSFET Models for VLSI Circuit Simulation: Theory and Practice*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1993.

[74] J. M. Rabaey, *Digital Integrated Circuits - a Design Perspective*. 1996.

[75] P. Yu, S. X. Shi, and D. Z. Pan, "Process variation aware OPC with variational lithography modeling," in *Design Automation Conference*, pp. 785–790, 2006.

[76] R. Singhal, A. Balijepalli, A. Subramaniam, F. Liu, S. Nassif, and Y. Cao, "Modeling and analysis of non-rectangular gate for post-lithography circuit simulation," in *DAC '07: Proceedings of the 44th annual Design Automation Conference*, (New York, NY, USA), pp. 823–828, ACM, 2007.

[77] H. H. Chen and D. D. Ling, "Power supply noise analysis methodology for deep-submicron VLSI chip design," in *Design Automation Conference*, pp. 638–643, 1997.

[78] J. Cong, Z. Pan, L. He, C.-K. Koh, and K.-Y. Khoo, "Interconnect design for deep submicron ics," in *ICCAD '97: Proceedings of the 1997 IEEE/ACM international conference on Computer-aided design*, (Washington, DC, USA), pp. 478–485, IEEE Computer Society, 1997.

[79] W. C. Elmore, "The transient response of damped linear networks with particular regard to wideband amplifiers," *Journal of Applied Physics*, vol. 19, no. 1, p. 55, 1948.

[80] P. Penfield, Jr. and J. Rubinstein, "Signal delay in RC tree networks," in *DAC '81: Proceedings of the 18th conference on Design automation*, (Piscataway, NJ, USA), pp. 613–617, IEEE Press, 1981.

[81] L. T. Pillage, X. Huang, and R. A. Rohrer, "AWEsim: asymptotic waveform evaluation for timing analysis," in *DAC '89: Proceedings of the 26th ACM/IEEE conference on Design automation*, (New York, NY, USA), pp. 634–637, ACM, 1989.

[82] L. Pillage and R. Rohrer, "Asymptotic waveform evaluation for timing analysis," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 9, pp. 352–366, Apr 1990.

[83] Q. Yu and E. S. Kuh, "Exact moment matching model of transmission lines and application to interconnect delay estimation," *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 3, no. 2, pp. 311–322, 1995.

[84] D. Edelstein, J. Heidenreich, R. Goldblatt, W. Cote, C. Uzoh, N. Lustig, P. Roper, T. McDevitt, W. Motsiff, A. Simon, J. Dukovic, R. Wachnik, H. Rathore, R. Schulz, L. Su, S. Luce, and J. Slattery, "Full copper wiring in a sub-0.25 μm CMOS ULSI technology," in *Electron Devices Meeting, 1997. IEDM '97. Technical Digest., International*, pp. 773–776, Dec 1997.

[85] A. Naeemi, R. Sarvari, and J. Meindl, "Performance comparison between carbon nanotube and copper interconnects for gigascale integration (GSI)," *Electron Device Letters, IEEE*, vol. 26, pp. 84–86, Feb. 2005.

[86] N. Srivastava and K. Banerjee, "Performance analysis of carbon nanotube interconnects for VLSI applications," in *Computer-Aided Design,*

2005. ICCAD-2005. IEEE/ACM International Conference on, pp. 383–390, Nov. 2005.

[87] J. Li, Q. Ye, A. Cassell, H. T. Ng, R. Stevens, J. Han, and M. Meyyappan, "Bottom-up approach for carbon nanotube interconnects," *Applied Physics Letters*, vol. 82, no. 15, pp. 2491–2493, 2003.

[88] J. P. Fishburn and C. A. Schevon, "Shaping a distributed-RC line to minimize Elmore delay," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 42, no. 12, p. 1020, 1995.

[89] C.-P. Chen, Y.-P. Chen, and D. F. Wong, "Optimal wire-sizing formula under the Elmore delay model," in *DAC*, p. 487, 1996.

[90] J. P. Fishburn, "Shaping a VLSI wire to minimize Elmore delay," in *European Design and Test Conference*, pp. 244–251, 1997.

[91] C.-P. Chen and D. F. Wong, "Optimal wire-sizing function with fringing capacitance consideration," in *DAC*, p. 604, 1997.

[92] Y. Gao and D. F. Wong, "Shaping a VLSI wire to minimize delay using transmission line model," in *ICCAD*, p. 611, 1998.

[93] J. Cong and Z. Pan, "Wire width planning for interconnect performance optimization," *TCAD*, vol. 21, no. 3, p. 319, 2002. 0278-0070.

[94] J. Cong and K. shing Leung, "Optimal wiresizing under the distributed Elmore delay model," in *Proc. Int. Conf. on Computer Aided Design*, pp. 634–639, 1993.

[95] C. C. N. Chu and M. D. F. Wong, "Greedy wire-sizing is linear time," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 18, no. 4, p. 398, 1999. 0278-0070.

[96] J. Cong and Z. Pan, "Interconnect performance estimation models for design planning," *TCAD*, vol. 20, no. 6, p. 739, 2001. 0278-0070.

[97] "*http://www.sciencetimeline.net/1651.htm*."

[98] S. X. Shi and D. Z. Pan, "Wire sizing with scattering effect for nanoscale interconnection," in *ASP-DAC '06: Proceedings of the 2006 conference on Asia South Pacific design automation*, (Piscataway, NJ, USA), pp. 503–508, IEEE Press, 2006.

[99] K. Fuchs, "The conductivity of thin metallic films according to the electron theory of metals," in *Cambridge Philosophical Society*, vol. 34, pp. 100–108, 1938.

[100] E. Sondheimer, "The mean free path of electrons in metals," *Adv. Phys.*, vol. 1, pp. 1–42, 1952.

[101] A. F. Mayadas and M. Shatzkes, "Electrical-resistivity model for polycrystalline films: the case of arbitrary reflection at external surfaces," *Phys. Rev*, vol. B 1, no. 4-15, pp. 1382–1389, 1970.

[102] Z. Tesanovic, M. V. Jaric, and S. Maekawa, "Quantum transport and surface scattering," *Phys. Rev. Lett.*, vol. 57, pp. 2760–2763, 1986.

[103] R. Dannenberg and A. H. King, "Behavior of grain boundary resistivity in metals predicted by a two-dimensional model," *Journal of Applied Physics*, vol. 88, no. 5, p. 2623, 2000.

[104] C. Durkan and M. E. Welland, "Size effects in the electrical resistivity of polycrystalline nanowires," *Phys. Rev. B*, vol. 61, p. 1421514218, 2000.

[105] J. Vancea, G. Reiss, and H. Hoffmann, "Mean-free-path concept in polycrystalline metals," *Phys. Rev. B*, vol. 35, pp. 6435–6437, 1987. MS/FS models have been extensively tested against experimental data for thin films in this paper.

[106] J. Vancea, "Unconventional features of free electrons in polycrystalline metal films," *International Journal of Modern Physics B*, vol. 3, no. 10, pp. 1455–1501, 1989.

[107] G. Steinlesberger, M. Engelhardt, G. Schindler, W. Steinhogl, A. v. Glasow, K. Mosig, and E. Bertagnolli, "Electrical assessment of copper damascene interconnects down to sub-50nm feature sizes," *Microelectron. Eng.*, vol. 64, no. 1, pp. 409–416, 2002. 604974.

[108] W. Wen and K. Maex, "Studies on size effect of copper interconnect lines," in *International Conference on Solid-State and Integrated-Circuit Technology (ICSICT)*, vol. 1, p. 416, 2001.

[109] C.-U. Kim, J. Park, N. Michael, P. Gillespie, and R. Augur, "Study of electron-scattering mechanism in nanoscale Cu interconnects," *Journal*

of Electronic Materials, vol. 32, no. 10, pp. 982–987(6), 2003.

[110] "International technology roadmap for semiconductors (ITRS)," 2004.

[111] A. Raychowdhury and K. Roy, "A circuit model for carbon nanotube interconnects: comparative study with Cu interconnects for scaled technologies," in *ICCAD*, p. 237, 2004.

[112] N. Srivastava and K. Banerjee, "A comparative scaling analysis of metallic and carbon nanotube interconnections for nanometer scale VLSI technologies," in *the 21st International VLSI Multilevel Interconnect Conference (VMIC)*, (Waikoloa, HI), pp. 393–398, 2004.

[113] P. Kapur, J. McVittie, and K. Saraswat, "Technology and reliability constrained future copper interconnects-part I: Resistance modeling," *IEEE Transactions on Electron Devices*, vol. 49, no. 4, pp. 590–597, 2002.

[114] S. H. Brongersma, K. Vanstreels, W. Wu, W. Zhang, D. Ernur, J. D'Haen, V. Terzieva, M. Van Hove, T. Clarysse, L. Carbonell, W. Vandervorst, W. De Ceuninck, and K. Maex, "Copper grain growth in reduced dimensions," in *Interconnect Technology Conference*, pp. 48–50, 2004.

[115] J. Rubinstein, P. Penfield, and M. A. Horowitz, "Signal delay in RC tree networks," *TCAD*, vol. 2, no. 3, p. 202, 1983. 0278-0070.

[116] R. Gupta, B. Krauter, B. Tutuianu, J. Willis, and L. T. Pileggi, "The Elmore delay as bound for RC trees with generalized input signals," in

DAC '95: Proceedings of the 32nd ACM/IEEE conference on Design automation, (New York, NY, USA), pp. 364–369, ACM, 1995.

[117] S. Borkar, "Design challenges of technology scaling," *IEEE Micro*, vol. 19, no. 4, pp. 23–29, 1999. 0272-1732.

[118] L. Euler, *Methodus inveniendi lineas curvas maximi minimive proprietate gaudentes sive solutio pro blematis isoperimetrici lattissimo sensu accepti*. 1744.

[119] H. Chang and S. Sapatnekar, "Statistical timing analysis under spatial correlations," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 24, pp. 1467–1482, Sept. 2005.

[120] M. Orshansky and A. Bandyopadhyay, "Fast statistical timing analysis handling arbitrary delay correlations," in *Design Automation Conference, 2004. Proceedings. 41st*, pp. 337–342, 2004.

[121] C. Visweswariah, K. Ravindran, K. Kalafala, S. G. Walker, S. Narayan, D. K. Beece, J. Piaget, N. Venkateswaran, and J. G. Hemmett, "First-order incremental block-based statistical timing analysis," *TCAD*, vol. 25, no. 10, p. 2180, 2006.

[122] A. Agarwal, D. Blaauw, and V. Zolotov, "Statistical timing analysis for intra-die process variations with spatial correlations," in *ICCAD*, p. 907, 2003.

[123] Y. Zhan, A. J. Strojwas, X. Li, L. T. Pileggi, D. Newmark, and M. Sharma, "Correlation-aware statistical timing analysis with non-gaussian delay distributions," in *DAC*, pp. 77–82, 2005.

[124] M. Mani, A. Devgan, and M. Orshansky, "An efficient algorithm for statistical minimization of total power under timing yield constraints," in *DAC '05: Proceedings of the 42nd annual Design Automation Conference*, (New York, NY, USA), pp. 309–314, ACM, 2005.

[125] R. Chen, L. Zhang, V. Zolotov, C. Visweswariah, and J. Xiong, "Static timing: back to our roots," in *ASP-DAC '08: Proceedings of the 2008 Asia and South Pacific Design Automation Conference*, (Los Alamitos, CA, USA), pp. 310–315, IEEE Computer Society Press, 2008.

[126] G. Gerosa, S. Gary, C. Dietz, P. Dac, K. Hoover, J. Alvarez, H. Sanchez, P. Ippolito, N. Tai, S. Litch, J. Eno, J. Golab, N. Vanderschaaf, and J. Kahle, "A 2.2W, 80MHz superscalar RISC microprocessor," *IEEE Journal of Solid-State Circuits*, vol. 29, no. 12, pp. 1440–1454, 1994.

[127] R. Chen and H. Zhou, "Statistical timing verification for transparently latched circuits," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 25, pp. 1847–1855, Sept. 2006.

[128] M. C.-T. Chao, L.-C. Wang, K.-T. Cheng, and S. Kundu, "Static statistical timing analysis for latch-based pipeline designs," in *ICCAD '04: Proceedings of the 2004 IEEE/ACM International conference on Computer-*

aided design, (Washington, DC, USA), pp. 468–472, IEEE Computer Society, 2004.

[129] L. Zhang, Y. Hu, and C. C. Chen, "Statistical timing analysis in sequential circuit for on-chip global interconnect pipelining," in *DAC*, pp. 904–907, 2004.

[130] J.-f. Lee, D. T. Tang, and C. K. Wong, "A timing analysis algorithm for circuits with level-sensitive latches," in *ICCAD '94: Proceedings of the 1994 IEEE/ACM international conference on Computer-aided design*, (Los Alamitos, CA, USA), pp. 743–748, IEEE Computer Society Press, 1994.

[131] S. Srivastava and J. Roychowdhury, "Interdependent latch setup/hold time characterization via Euler-Newton curve tracing on state-transition equations," in *DAC '07: Proceedings of the 44th annual Design Automation Conference*, (New York, NY, USA), pp. 136–141, ACM, 2007.

[132] T. Karnik, B. Bloechel, K. Soumyanath, V. De, and S. Borkar, "Scaling trends of cosmic ray induced soft errors in static latches beyond 0.18 μ," in *VLSI Circuits, 2001. Digest of Technical Papers. 2001 Symposium on*, pp. 61–62, 2001.

[133] A. Components, "TSMC 0.18μm process 1.8-Volt SAGE-X standard cell library databook," release 4.0,, Feb. 2002.

[134] N. H. Weste and D. Harris, *CMOS VLSI Design: A Circuits and Systems Perspective*. Pearson Higher Education, 2004. chapter 7.4 static sequencing element methodology pp.425.

[135] B. Zhang, A. Arapostathis, S. Nassif, and M. Orshansky, "Analytical modeling of SRAM dynamic stability," in *ICCAD*, p. 315, 2006.

[136] Z. Vukic, L. Kuljaca, D. Donlagic, and S. Tesnjak, *Nonlinear Control Systems*. Marcel Dekker Inc., 2003.

[137] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, "Parameter variations and impact on circuits and microarchitecture," *Design Automation Conference*, pp. 338–432, 2003.

[138] A. Asenov, "Random dopant induced threshold voltage lowering and fluctuations in sub-0.1 μm MOSFET's: a 3-D "atomistic" simulation study," *Electron Devices, IEEE Transactions on*, vol. 45, pp. 2505–2513, Dec 1998.

[139] S. R. Nassif, "Design for variability in DSM technologies," in *ISQED '00: Proceedings of the 1st International Symposium on Quality of Electronic Design*, (Washington, DC, USA), p. 451, IEEE Computer Society, 2000.

[140] F. Brglez, D. Bryan, and K. Kozminski, "Combinational profiles of sequential benchmark circuits," in *Circuits and Systems, 1989., IEEE International Symposium on*, pp. 1929–1934 vol.3, May 1989.

Vita

Xiaokang (Sean) Shi was born in Beijing, China, son of Bingkang Shi and Xiaodong Li. He received his Bachelor of Science degree in Physics, Master of Science degree in Microelectronics from Peking University and Master of Science in Engineering in Computer Engineering from The University of Texas at Austin in 2001, 2004 and 2008 respectively. In spring of 2005, he joined Design Automation Lab in ECE department. He has published about 6 journal papers and 20 conference papers with 3 patents issued and 1 patent filed. His research interests cover modeling and optimization for different kinds of problems in process simulation (TCAD), device modeling, physical design, timing analysis, reliability verification and lithography. He has been awarded IBM PhD Scholarship, Intel Scholarship on Information Science and Technology, Yang Fuqing & Wang Yangyuan Academicians Scholarship, Honor of Innovation in Peking University, Excellent Winner of "Science Academic Top 10" in Peking University, DAC Young Student Support Program Award, etc.

Permanent address: 2501 Lake Austin Blvd #J104
Austin, Texas 78703

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.